Junio, 2009

Análisis en Componentes de Imágenes Funcionales para la Ayuda al Diagnóstico de la Enfermedad del Alzheimer

Tesis Doctoral

Ignacio Alvarez Illán

UNIVERSIDAD DE GRANADA

Editor: Editorial de la Universidad de Granada
Autor: Ignacio Alvarez Illán
D.L.: GR. 3197-2009
ISBN: 978-84-692-5112-6

***Dr. Dr. Juan Manuel Górriz Saez**, profesor titular de la Universidad de Granada del Departamento de Teoría de la Señal, Telemática y Comunicaciones, **Dr. Javier Ramírez Perez de Inestrosa**, profesor titular de la Universidad de Granada del Departamento de Teoría de la Señal, Telemática y Comunicaciones, **Dr. Carlos Garcia Puntonet**, catedrático de la Universidad de Granada del Departamento de Arquitectura y Tecnología de Computadores*

CERTIFICAN

*Que la memoria titulada: '**Análisis en Componentes de Imágenes Funcionales para la Ayuda al Diagnóstico de la Enfermedad del Alzheimer**' ha sido realizado por **D. Ignacio Alvarez Illán** bajo nuestra dirección en el departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada, para optar al grado de doctor por la Universidad de Granada.*

Granada, 8 de Junio de 2009

Fdo.: J.M. Górriz *Fdo.: J. Ramírez*

Fdo.: C.G. Puntonet

Prefacio

La presente tesis presenta varias técnicas nóveles para la ayuda al diagnóstico del Alzheimer por computador, que han dado resultados prometedores dentro del campo, y cuya aplicación se ha testado con resultados muy satisfactorios. El trabajo multidisciplinar que supone el desarrollo de esta herramienta, ha sido avalado por su publicación en revistas indexadas por el ISI englobadas en diversas áreas del conocimiento, asi como presentado en congresos internacionales de prestigio. Además, este trabajo formará parte de un proyecto más amplio en desarrollo para crear una herramienta informática completa que esté disponible en la práctica clínica de los hospitales, en colaboración con las empresas PET-Cartuja y PTEC. La cooperación del grupo de investigación bajo el que opera este trabajo con el proyecto ADNI (Alzheimer Disease NeuroImaging Initiative), hace que el proyecto tome un carácter internacional y se sirva de una de las bases de datos más amplia a nivel mundial en lo que se refiere a la enfermedad del Alzheimer. La importancia del diagnóstico de esta enfermedad en una sociedad en la que la edad media de la población es cada vez mayor, hace que una contribución en esta dirección sea necesaria para buscar tratamientos eficaces.

Parte del trabajo realizado en esta tesis ha derivado en las siguientes publicaciones en revistas indexadas por el ISI:

I. I. Alvarez Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, M- López, C. G. Puntonet, F. Segovia. Alzheimer's Diagnosis Using Eigenbrains and Support Vector Machines. IET Electronics Letters 45 (2009) 7 342-343

II. D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. A. Illán, C. G. Puntonet, M. Gómez-Río. Analysis of SPECT brain images for the diagnosis of Alzheimer's disease using moments and support vector machines. Accepted in Neuroscience Letters.

III. J. Ramírez, J. M. Górriz, R. Chaves, D. Salas-Gonzalez, I. Álvarez, M. López and F. Segovia. SPECT Image Classification Using Random Forests. IET Electronics Letters. Volume 45 (2009), Issue 12, p. 604-605.

IV. J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, M. López, I. Álvarez, M. Gómez-Río, Computer Aided Diagnosis of Alzheimer Type Dementia Combining Support Vector Machines and Discriminant Set of Features. Accepted in Information Sciences.

V. M. López, J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, C. G. Puntonet. Automatic Tool for the Alzheimer's Disease Diagnosis Using PCA and Bayesian Classification Rules. IET Electronics Letters. Vol 45 (2009), no. 8. pp. 389-391.

Asimismo ha producido las siguientes aportaciones a congresos internacionales.

I. Autores: J.M. Gorriz, J. Ramirez, A. Lassl, D. Salas-gonzález, E. Lang, C.G. Puntonet, I.A. Illán, M. Lopez, M. Gomez-Rio Título: Automatic computer aided diagnosis tool using component based SVM Tipo de participación: Póster Congreso: IEEE Nuclear Science symposium conference Publicación: IEEE NSS-MIC Record, pags 4392 - 4395 Lugar celebración: Dresden, Alemania Fecha: 19 - 25 Octubre 2008

II. Autores: I.A.Illan, M.Lopez, J. M. Gorriz, J. Ramirez, C.G. Puntonet, Diego salas gonzalez Título: Automatic classificatio system for the diagnosis of Alzheimer using component based SVM agregations Tipo de

participación: Poster Congreso: Proceedings of the 15th International conference on neural information processing (ICONIP) Publicación: Springer LNCS volumes. Lugar celebración: Auckland, Nueva Zelanda Fecha: 25-28 noviembre 2008

III. Autores: D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. Álvarez, F. Segovia, C. G. Puntonet. Título: Analysis of Brain SPECT Images for the diagnosis of Alzheimer disease using first and second order moments. Tipo de participación: Poster Congreso: The International Work-conference on the Interplay between Natural and Artificial Computation (IWINAC) Publicación: Springer LNCS volumes. Lugar celebración: Santiago de Compostela (Spain) Fecha: June 2009

IV. Autores: M. López, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, M. Gómez-Río. Título: Support Vector Machines and Neural Networks for the Alzheimer's Disease Diagnosis Using PCA. Tipo de participación: Poster Congreso: The International Work-conference on the Interplay between Natural and Artificial Computation (IWINAC) Publicación: Springer LNCS volumes. Lugar celebración: Santiago de Compostela (Spain) Fecha: June 2009

V. Autores: J. Ramírez, R. Chaves, J. M. Górriz, I. Álvarez, M. López, D. Salas-Gonzalez, F. Segovia. Título: Functional Brain Image Classification Techniques for Early Alzheimer Disease Diagnosis. Tipo de participación: Poster Congreso: The International Work-conference on the Interplay between Natural and Artificial Computation (IWINAC) Publicación: Springer LNCS volumes. Lugar celebración: Santiago de Compostela (Spain) Fecha: June 2009

VI. Autores: J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, F. Segovia, C. G. Puntonet. Título: Classification of SPECT images using clustering techniques revisited Tipo de participación: Poster Congreso: The International Work-conference on the Interplay between Natural and Artificial Computation (IWINAC) Publicación: Springer LNCS volumes. Lugar celebración: Santiago de Compostela (Spain) Fecha: June 2009

VII. Autores: D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. Álvarez, F. Segovia, C. G. Puntonet. Título: Selecting regions of interest

for the diagnosis of Alzheimer's disease in brain SPECT images using Welch's t-test. Tipo de participación: Poster Congreso: The 10th International Work-Conference on Artificial Neural Networks, (IWANN). Publicación: Springer LNCS volumes. Lugar celebración: Salamanca (Spain) Fecha: June 2009

viii. Autores: M. López, J. Ramírez, J.M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia and C.G. Puntonet. Título: Automatic System for Alzheimer's Disease Diagnosis Using Eigenbrains and Bayesian Classification Rules Tipo de participación: Poster Congreso: The 10th International Work-Conference on Artificial Neural Networks, (IWANN). Publicación: Springer LNCS volumes. Lugar celebración: Salamanca (Spain) Fecha: June 2009

ix. Autores: J. Ramírez, R. Chaves, J. M. Górriz, M. López, D. Salas-Gonzalez, I. Álvarez, F. Segovia. Título: SPECT image classification techniques for computer aided diagnosis of the Alzheimer disease. Tipo de participación: Poster Congreso: The 10th International Work-Conference on Artificial Neural Networks, (IWANN). Publicación: Springer LNCS volumes. Lugar celebración: Salamanca (Spain) Fecha: June 2009

x. Autores: I. Álvarez, J. M. Górriz, J. Ramírez, M. López, D. Salas-Gonzalez, C. G. Puntonet, F. Segovia, B. Prieto. Título: Alzheimer's Diagnosis Using Eigenbrains and Support Vector Machines. Tipo de participación: Poster Congreso: The 10th International Work-Conference on Artificial Neural Networks, (IWANN). Publicación: Springer LNCS volumes. Lugar celebración: Salamanca (Spain) Fecha: June 2009

Por otro lado, parte del trabajo que se describe en esta tesis ha sido enviado a las siguientes revistas indexadas por el ISI;

i. I. Alvarez-Illan, J. M. Gorriz, D. Salas-Gonzalez, J. Ramirez, M.Lopez, F. Segovia, C.G. Puntonet. Independent Component Analysis of SPECT Images to assist the Alzheimer's Disease Diagnosis. Enviado a Pattern Recognition letters.

ii. I. Alvarez-Illan, J. M. Gorriz, D. Salas-Gonzalez, J. Ramirez, M.Lopez, F. Segovia, C.G. Puntonet. M. Gomez-Rio, and the Alzheimer's Disease

Neuroimaging Initiative. ^{18}F-FDG PET Imaging For Computer Aided Alzheimer's Diagnosis. Enviado a Neuroimage.

III. I. Alvarez-Illan, J. M. Gorriz, D. Salas-Gonzalez, J. Ramirez, M.Lopez, F. Segovia, C.G. Puntonet.Pasting votes of component based SVM classiffiers for the diagnosis of Alzheimer's disease. Enviado a Applied Soft Computing.

estando pendiente de resolución.

El presente trabajo esta estructurado en tres partes. En la primera parte introducimos los antecedentes del trabajo, así como la tecnología en la que están basadas las aportaciones en el campo del aprendizaje estadístico que presenta esta investigación. Primeramente comenzaremos describiendo la patología bajo estudio, la enfermedad de Alzheimer, los efectos de esta enfermedad en la sociedad actual así como la importancia del diagnóstico precoz de la misma. A continuación describiremos las técnicas de tomografía que nos proporcionan los "mapas" cerebrales de activación, claves para el diagnóstico mediante computador en los llamados sistemas de ayuda al diagnóstico usando computadores (del inglés "computer aided diagnosis -CAD- systems"). Finalmente describiremos el estado del arte en el campo de las neurociencias para el diagnóstico y/o evaluación cuantitativa de imágenes cerebrales funcionales de manera automática.

En la segunda parte se describen las bases de datos sobre las que se trabajará, asi como el procesado que se llevará a cabo antes de comenzar la tarea de clasificación. En la última parte se describen tres novedosas técnicas para la ayuda al diagnóstico, basadas en análisis de componentes, componentes principales (PCA) y componentes independientes (ICA).

Índice general

I Fundamentos **1**

1. Introducción **3**

1.1. La enfermedad de Alzheimer 5

1.1.1. Técnicas de ayuda al diagnóstico del Alzheimer 8

1.2. Técnicas Tomograficas empleadas en medicina nuclear 11

1.2.1. SPECT 12

1.2.2. PET 14

2. CAD **15**

3. SPM **19**

3.1. Preprocesado en SPM 20

3.2. Análisis Estadístico 21

4. Aprendizeje estadístico **25**

4.1. Métodos Kernel 29

4.2. Clasificadores estadísticos 33

4.2.1. Nearest Mean 34

4.2.2. Clasificador linear de Fisher 35

4.3. Selección del subconjunto de características 36

4.3.1. Métodos de filtro 39

4.3.2. Métodos de Envoltura 44

4.4. Parametros de valoración del rendimiento de un clasificador . 45

4.4.1. Curva ROC 47

4.5. Métodos de Validación Cruzada 48

4.5.1. Validación por sub-muestreo aleatorio repetido 49

4.5.2. Validación Cruzada K-pliegues 50

4.5.3. Validación dejar uno fuera 50

5. SVM **51**

5.1. SVM Lineal 52

5.1.1. Clases linealmente separables 52

5.1.2. Clases linealmente no separables 56

5.2. SVM No Lineal 60

5.3. Agregado de SVM 62

II Desarrollos Experimentales 65

6. Bases de Datos 67

6.1. ADNI 68

6.1.1. Protocolo de adquisición 68

6.1.2. Criterios de Etiquetado 71

6.2. Virgen de las Nieves 72

6.3. PET Cartuja 73

7. Preprocesado 75

7.1. Adquisición de Imágenes 76

7.2. Reconstrucción de Imágenes 76

7.3. Registro de Imágenes 78

7.4. Normalización de intensidad 80

III Ánalisis en componentes 83

8. Componentes 85

8.1. Método de Componentes 86

8.2. Extracción de características basada en Factorización 87

8.3. Agregación del Conjunto de Decisiones basadas en SVM . . . 89

8.3.1. Voto por mayoría 90

8.3.2. Voto por Relevancia 91

8.4. Experimientos 94

8.5. Resultados 94

9. PCA **99**

9.1. Análisis de Componentes Principales 100

9.2. Reducción de la dimensionalidad mediante selección de CPs . 102

9.2.1. Eigenbrains 104

9.2.2. Selección de características a partir de CPs 107

9.2.3. Selección mediante el criterio de Fisher 109

9.3. Filtro de Componentes Principales 109

9.4. Eliminación de Correlación mediante CPs 110

9.5. Experimentos 114

9.5.1. PCA 114

9.5.2. PCA como proyección 116

9.6. Resultados 116

9.6.1. PCA 116

9.6.2. PCA como proyección 118

10.ICA **123**

10.1. Análisis de Componentes Independientes 124

10.1.1. Información mutua 125

10.1.2. Función de contraste 126

10.1.3. Algoritmos para ICA 127

10.2. Reduccion de la dimensionalidad mediante ICs 131

10.3. ICA aplicado a imágenes representativas 132

10.4. Experimentos 134

10.4.1. Experimentos con VN 134

10.4.2. Experimentos con ADNI 136
10.5. Resultados 137
10.5.1. Resultados con VN 137
10.5.2. Resultados con ADNI 138

11. Conclusiones **141**
11.1. Trabajo futuro 145

Referencias **148**

Índice de figuras

1.1. Gamma Cámara Picker Prism 3000 13

3.1. Resultados de SPM para clasificación con el modelo descrito en el texto 24

4.1. Comparación del argoritmo de clasificación basado en media mas cercana y fisher lineal. 35

4.2. Probabilidad de agrupar n puntos en un espacio de características m-dimensional en 2 clases linealmente separables. . . 37

4.3. Secciones transversales de: *Columna izquierda*: Un paciente normal. *Columna central*: Un paciente DTA. *Columna derecha*: Máscara 41

4.4. Representación en el espacio ROC 48

5.1. Mapeo del espacio de características a otro espacio donde la frontera de separación se hace lineal 60

7.1. Tres imágenes SPECT. *Columna izquierda*: Imagen original. *Columna central*: Plantilla. *Columna derecha*: Imagen transformada tras el proceso de registro. 80

7.2. 20 histogramas de intensidad de diferentes imágenes de la base de datos SPECT 81

8.1. Cortes sagital, coronal y transversal de una imagen SPECT cerebral con la componente de la imagen remarcada 89

8.2. Mapa de las ROIs de un conjunto de imágenes SPECT obtenidas mediante el valor de A_i 95

8.3. Mapa de las regiones de interés según su precisión considerando solo pacientes ATD-1 (izquierda) o considerando solo pacientes ATD-2 y ATD-3 (derecha) 95

8.4. Precisión para del sistema basado en SVM con umbral T variable y para distintos factores de reducción. 97

9.1. Varianza explicada por los 76 primeros eigenbrains 107

9.2. Representación en términos de Eigenbrains de la imagen p con media cero. Se muestran tres cortes transaxiales respresentativos. 115

9.3. Precisión de dos clasificadores usando diferentes kernels frente al numero de eigenbrains usado para contruir los vectores de características para los metodos de seleccionar los eigenbrains. 117

9.4. Precisión del clasificador para diferentes factores de compresión $f = \{1/2^3, 1/4^3, 1/6^3, 1/8^3\}$ y diferentes kernels frente al número de eigenbrains usado. 118

9.5. Clasificador lineal y vectores de entrenamiento $\mathbf{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio de vectores característicos (izquierda, cortes transaxiales) 120

9.6. Clasificador lineal y vectores de entrenamiento $\mathtt{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio PCA de eigenbrains (izquierda, cortes transaxiales) 121

10.1. Proceso de separación de las 4 imagenes representativas originales en 4 fuentes independientes. Se muestran tres cortes transaxiales del cerebro. 135

10.2. Representación de un paciente de Alzheimer en términos de los coeficientes de ICA siguiendo el método I. 136

10.3. Clasificador lineal y vectores de entrenamiento $\mathtt{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio de ICs (izquierda, cortes transaxiales) 138

11.1. Representación en el espacio ROC de los resultados de clasificación de los mismos pacientes usando las técnicas de proyección, PCA e ICA para diferentes trazadores y kernels. Se tienen en cuenta solo los pacientes AD y los NC 146

11.2. Representación en el espacio ROC de los resultados de clasificación de los mismos pacientes usando las técnicas de proyección, PCA e ICA para diferentes trazadores y kernels. Se tienen en cuenta todos los pacientes AD, MCI y NC 147

Índice de tablas

4.1. Posibles resultados del test en función de la etiqueta 46

8.1. Medidas estadísticas del rendimiento de la prueba por voto en mayoría, con un número fijo de vóxeles contenidos en cada componente ó con un número fijo de componentes. VAF (Stoeckel et˜al., 2001) se muestra como referencia. 96

8.2. Medidas estadísticas para el voto por relevancia, véase el texto para una mayor descripción. 96

9.1. Medidas del rendimiento del metodo de PCA para los tres grupos muestrales, en referencia a la simple proyección. 119

10.1. Medidas de precisión del método de ICA 138

10.2. Medidas del rendimiento del metodo de ICA para los tres grupos muestrales, en referencia a la simple proyección 139

11.1. Comparación de las medidas de precisión de los métodos de ICA y PCA, para los 3 grupos. 144

Parte I

Fundamentos

CAPÍTULO 1

Introducción

La enfermedad de Alzheimer es la forma más común de demencia entre las personas mayores. El Alzheimer es una enfermedad neurodegenerativa que afecta a la memoria en su primer estadio para ir gradualmente deteriorando todas las funciones cognitivas asi como el comportamiento y finalmente causando la muerte. Es una patología que va asociada a la edad llegándose a incrementar progresivamente su prevalencia hasta alcanzar aproximadamente el 30 % de los mayores de 85 años. De acuerdo con la Fundación Alzheimer España, cerca de 700.000 personas sufren la enfermedad en España y se manifiestan más de 100.000 nuevos enfermos cada año. Más de dos millones de familiares de pacientes de Alzheimer ven su vida trastornada como consecuencia de la enfermedad. Según las últimas estimaciones, 8 millones de personas sufren la enfermedad de Alzheimer en Europa alcanzándose la cifra de 30 millones en todo el mundo. Teniendo en cuenta el envejecimiento de la población y el futuro incremento de personas mayores de 80 años, se prevé que el número de enfermos se duplique en 2020 y triplique en 2050 Carr et˜al.

(1997). Es uno de los retos más importantes a los que deberá enfrentarse la sociedad en el transcurso del siglo XXI.

En la práctica clínica, el diagnóstico esta basado en cuidadosos analisis clínico, entrevistas con el paciente y sus familiares, y test neuropsicologicos. A menudo, las imagenes funcionales del cerebro sirven de ayuda para el diagnóstico, aunque la enfermedad solo podrá ser confirmada tras la autopsia. En la actualidad el diagnóstico del Alzheimer representa todavía un reto, especialmente en su etapa mas temprana, cuando hay mas oportunidades de tratar sus síntomas así como de probar y desarrollar nuevos tratamientos.

El descubrimiento de nuevos fármacos efectivos para el tratamiento de los síntomas de la enfermedad junto con otros agentes que están siendo clínicamente evaluados, abriría una nueva esperanza en el tratamiento de la enfermedad de Alzheimer. Existe un claro consenso en la necesidad de desarrollar técnicas de diagnóstico precoz más efectivas para una rápida intervención, para prevenir o disminuir la progresión de la enfermedad y para que se pueda hacer el máximo aprovechamiento de los servicios y tratamientos disponibles para los pacientes que la sufren. El diagnóstico en las primeras fases de la enfermedad ayuda a los pacientes y a sus familias a planear el futuro, les da tiempo para discutir las opciones de cuidado mientras el paciente todavía puede participar en la toma de decisiones y, no menos importante, ofrece la mejor oportunidad para tratar los síntomas de la enfermedad. Ningún tratamiento puede detener la enfermedad de Alzheimer. Sin embargo, en algunas personas, en las fases precoz y media de la enfermedad, ciertos medicamentos pueden prevenir el empeoramiento de algunos síntomas durante un tiempo limitado.

Varias líneas convergentes de investigación sugieren que el proceso neurodegenerativo asociado con la demencia comienza varios años antes de que las características clínicas puedan ser detectadas con los intrumentos actuales. La duración precisa de este preriodo preclínico y los detalles de los procesos moleculares que se generan son todavía desconocidos. Esta incertidumbre sobre las fases tempranas sin síntomas de la enfermedad reside en la ausencia de técnicas validadas y precisas para el diagnóstico.

Un numero elevado de estudios han examinado la capacidad de predicción de la medicina nuclear con respecto a la enfermedad del Alzheimer y otras demencias (Hoffman et˜al., 2000; Silverman et˜al., 2001; Higdon et˜al.,

2004). Las tomografías de emisión computerizadas ECT (Emission-computed tomography) han sido apliamente usadas en la investigación biomédica y en la práctica clínica durante las últimas tres décadas. ECT difiere de otras técnicas de imagen médicas como las resonancias magnéticas en que producen mapas de las funciones fisiológicas, en lugar de formar imagenes de estructuras anatómicas. Las imágenes tomográficas radiofarmaceuticas proporcionan mapas 3D *in vivo* de fármacos etiquetados mediante isótopos radioactivos, que son inyectados al paciente normalmente por via intravenosa. Habitualmente, estas imágenes son valoradas por expertos, quienes evaluan visualmente la presencia de características de la demencia. Sin embargo, la precisión con que actualmente se realiza el diagnóstico precoz de este tipo de enfermedades neurodegenerativas no supera el 70 % Cummings et˜al. (1998); Carr et˜al. (1997) y, en muchas ocasiones, éstas no reciben el tratamiento adecuado durante la fase inicial de la enfermedad.

Todavía resulta necesario introducir nuevas técnicas más eficientes de procesado, modelado y clasificación de imágenes, así como mejorar el rendimiento de las técnicas existentes. Sin embargo, los métodos de clasificación estadística no han sido usados ampliamente en este área, a pesar de haber producido notables resultados para el reconocimiento de patrones en otras areas. Esto puede deberse al hecho de que las imagenes representan grandes cantidades de datos, y la mayoría de los estudios cuentan con un número reducido de imágenes (generalmente <100) (Ishii et˜al., 2006; Stoeckel et˜al., 2001, 2004).

El objetivo de este trabajo es el diseño de nuevas técnicas de ayuda el diagnóstico basadas en el uso del aprendizaje estadístico supervisado por computador, con el propósito de obtener herramientas útiles para el diagnóstico que superen las limitaciones de las actuales, y mejorar la precisión en el diagnóstico de la enfermedad del Alzheimer, sobre todo en su etapa temprana.

1.1. La enfermedad de Alzheimer

Entre las enfermedades neurológicas en las que se centra la tarea de clasificación de este proyecto podemos destacar la enfermedad de Alzheimer. La demencia es un desorden cerebral que afecta seriamente la habilidad de una persona para llevar a cabo sus actividades diarias. La enfermedad de

Alzheimer (AD por sus siglas en inglés) es la forma más común de demencia entre las personas mayores. Involucra las partes del cerebro que controlan el pensamiento, la memoria y el lenguaje. Los científicos aprenden cada día más sobre esta enfermedad pero hasta ahora sus causas son desconocidas y no se conoce ninguna cura. Se sabe que sólo en EE.UU aproximadamente 4 millones de personas padecen de la enfermedad y esta se manifiesta habitualmente después de los 60 años aumentando el riesgo con la edad. Aunque las personas más jóvenes también pueden sufrirla, es mucho menos común entre ellas. Aproximadamente el 3 % de hombres y mujeres entre los 65 y los 74 años la tienen, y casi la mitad de los hombres y mujeres de 85 años en adelante pueden tenerla. Es importante notar sin embargo, que la enfermedad de Alzheimer no es una etapa normal del envejecimiento.

La enfermedad de Alzheimer es llamada así por el médico alemán Alois Alzheimer, quien en 1906 notó cambios en el tejido cerebral de una mujer que había muerto de una rara enfermedad mental. Alzheimer encontró aglutinaciones anormales (ahora llamadas 'placas de amiloide') y una masa confusa de fibras (ahora denominada 'enredo de neurofibrillas'). Actualmente, estas placas y enredos en el cerebro son considerados el sello característico de la enfermedad de Alzheimer. Los científicos también han encontrado otros cambios en el cerebro de las personas que tienen esta enfermedad. Hay una pérdida de células nerviosas en áreas del cerebro que son vitales para la memoria y otras habilidades mentales. También hay disminución en el nivel de las sustancias químicas responsables de la transmisión de los complejos mensajes entre las células nerviosas. Por causa del bloqueo de estos mensajes entre las células nerviosas se pueden alterar los procesos normales de pensamiento y de memoria.

Los científicos no entienden aún completamente qué causa la enfermedad de Alzheimer. Probablemente no hay una sola causa, sino que existen varios factores que afectan a cada persona de diferente manera. La edad es el factor de riesgo conocido más importante de la enfermedad de Alzheimer. El número de personas que la sufren se duplica cada 5 años entre las personas mayores de 65 años.

La enfermedad de Alzheimer comienza lentamente. Al principio el único síntoma puede ser tener pequeños olvidos. Las personas con la enfermedad de Alzheimer pueden tener problemas para recordar eventos recientes, actividades o los nombres de personas o cosas familiares. Las operaciones simples

matemáticas pueden volverse difíciles de resolver. Estas dificultades pueden causar estado de irritabilidad, pero normalmente no son tan serias como para causar alarma. Sin embargo, mientras la enfermedad progresa, los síntomas se notan más fácilmente y se acentúan tanto, que son motivo suficiente para que el enfermo de Alzheimer o los miembros de su familia, busquen ayuda médica. Por ejemplo, las personas en la etapa avanzada de la enfermedad de Alzheimer pueden olvidar la forma de hacer tareas sencillas, como cepillarse los dientes o peinarse. Ya no pueden pensar claramente y empiezan a tener problemas para hablar, entender, leer o escribir. Después, las personas con la enfermedad de Alzheimer pueden ponerse ansiosas o agresivas, o deambular fuera de casa. Eventualmente los pacientes necesitan un cuidado permanente.

Un diagnóstico temprano y exacto de la enfermedad de Alzheimer ayuda a los pacientes y a sus familias a planear el futuro. Les da tiempo para discutir las opciones de cuidado mientras el paciente todavía puede participar en la toma de decisiones. El diagnóstico temprano también ofrece la mejor oportunidad para tratar los síntomas de la enfermedad. Actualmente la única manera definitiva para diagnosticar la enfermedad de Alzheimer es investigar sobre la existencia de placas y enredos en el tejido cerebral. Para observar el tejido cerebral los médicos deben esperar hasta que se haga la autopsia, examen del cuerpo que se hace después de que la persona muere. Por consiguiente, los médicos deben hacer un diagnóstico de 'posible' o 'probable' enfermedad de Alzheimer basada en la experiencia. Exámenes del cerebro con escáner pueden permitir al médico observar fotografías del cerebro para detectar la presencia de anormalidades en él. Hasta ahora es todo el 'partido' que se está sacando a las nuevas técnicas de imágenes cerebrales a nivel médico en los servicios de medicina nuclear, quedando abierta la posible automatización del proceso, así como la clasificación-detección de estas alteraciones desde un punto de vista matemático.

La enfermedad de Alzheimer es una enfermedad lenta, que comienza con leves problemas de memoria y termina con un severo daño cerebral. El curso que toma la enfermedad y la velocidad en que ocurren los cambios varía de una persona a otra. En promedio, los pacientes con la enfermedad viven entre 8 y 10 años después de que se les ha diagnosticado, aunque esta puede durar hasta 20 años. Ningún tratamiento puede detener la enfermedad de Alzheimer. Sin embargo, en algunas personas, en las fases temprana y media de la enfermedad, medicamentos como el Tacrine (Cognex), Donepezil (Ari-

cept), Rivastigmine (Exelon) o Galantamine (Razadyne R° (anteriormente conocida como Reminyl R°)) pueden prevenir el empeoramiento de algunos síntomas durante un tiempo limitado.

1.1.1. Técnicas de ayuda al diagnóstico del Alzheimer

El diagnóstico de la demencia se basa fundamentalmente en la evaluación clínica y ésta requiere una exhaustiva evaluación de la función cognitiva, en concreto memoria, atención, percepción, lenguaje, praxias y gnosias. La evaluación neuropsicológica puede subdividirse en dos niveles de complejidad: Un primer nivel que consiste en el uso de test breves, estandarizados y sencillos como el Mini-Mental State Examination (MMSE), que permitan alcanzar el diagnóstico de demencia; y un segundo nivel de mayor complejidad en el se refina la evaluación de la severidad del deterioro, al tiempo que se establecen los dominios de la función cognitiva que se hallan afectados. Existen diferentes escalas que otorgan un valor estandarizado en función del grado de afectación funcional, como CDR (Clinical Dementia Rating) o GDS (Global Deterioration Scale). En general se intenta que estas escalas permitan clasificar la demencia según los criterios clínicos clásicos: demencia leve, moderada o severa.

Los cuestionarios o escalas han sido diseñados para cuantificar determinadas funciones cognitivas, es decir, no establecen un diagnóstico, sino que cuantifican la severidad de la alteración de determinadas áreas intelectuales, siendo particularmente valiosos para discriminar entre envejecimiento normal y demencias leves. El diagnóstico siempre ha de realizarse en base a la historia clínica y de acuerdo con los criterios establecidos al respecto; los cuestionarios representan sólo una ayuda en el proceso de valoración.

Examen del Estado Mental Mínimo (MMSE)

El MMSE es uno de los test de screening más utilizados. Es un test que tiene alta dependencia del lenguaje y consta de varios items relacionados con la atención. Cada item tiene una puntuación, llegando a un total de 30 puntos. En la práctica diaria una puntuación menor de 24 sugiere demencia, entre 23-21 una demencia leve, entre 20-11 una demencia moderada y menor

de 10 de una demencia severa. Para poder efectuar el MMSE es necesario que el paciente se encuentre vigil y lúcido. En la demencia por enfermedad de Alzheimer la tasa promedio anual de cambio en la puntuación del MMSE es de 2-5 puntos por año, por lo que el test muestra su utilidad para el seguimiento de pacientes dementes. El MMSE tiene baja sensibilidad para el diagnóstico de deterioro cognitivo leve, la demencia frontal-subcortical y el deficit focal cognitivo. Las características esenciales que se evalúan son:

- Capacidad de atención, concentración y memoria.
- Capacidad de abstracción (cálculo).
- Capacidad de lenguaje y percepción viso- espacial.
- Orientación espacio- tiempo.
- Capacidad para seguir instrucciones básicas.

Esta prueba proporciona un instrumento para detección de deterioro cognitivo que se puede realizar en poco tiempo. Según sus autores, esto era especialmente importante en la demencia, ya que el paciente rápidamente se cansaba, y por tanto dejaba de mostrarse colaborador.

Escala de Deterioro Global - GDS

La Escala de Deterioro Global (Global Deterioration Scale - GDS (Sheikh and Yesavage, 1986)) establece siete estadíos posibles:

- 1 = normal
- 2 = deterioro muy leve
- 3 = deterioro leve
- 4 = deterioro moderado
- 5 = deterioro moderadamente severo
- 6= deterioro severo

- 7 = deterioro muy severo

La escala define cada estadío en términos operacionales y en base a un deterioro supuestamente homogéneo. Sin embargo, dado que la secuencia de aparición de los síntomas es a menudo variable, se ha argumentado que la inclusión de un paciente en un estadío de acuerdo a un criterio rígido podría conducir a errores; no obstante se trata de una de las escalas más completas, simples y útiles para la estimación de la severidad de la demencia. El CAED (1997) sugirió la utilización de esta escala para la gradación del síndrome demencial de la enfermedad de Alzheimer acompañado del Instrumento de Evaluación Funcional para Enfermedad de Alzheimer, FAST (Functional Assessment tool for Alzheimer's disease).

Clasificación Clínica de la Demencia - CDR

La evaluación de las demencias que no son Enfermedad de Alzheimer se realiza a través de la Clasificación clínica de las Demencias (Clinical Dementia Rating - CDR (Morris, 1993)) que es más general. Su escala establece cinco estadíos posibles:

- 0 = normal
- 0,5 = cuestionable
- 1 = demencia leve
- 2 = demencia moderada
- 3 = demencia severa

La estimación se realiza en base al rendimiento del sujeto en seis modalidades de tipo cognitivo y funcional. Estas modalidades son: memoria, orientación, razonamiento, actividades sociolaborales, actividades recreativas (hobbies o pasatiempos), y cuidado personal.

Escala de Evaluación para la Enfermedad de Alzheimer (ADAS)

En 1984 con la aparición del ADAS (Alzheimer's Disease Assessment Scale) o Escala de Evaluación para la Enfermedad de Alzheimer (EEEA) fue posible contar con un instrumento fiable, breve, diseñado especialmente para la enfermedad de Alzheimer y capaz de medir puntualmente los síntomas característicos de la misma así como su progresión a estadios más avanzados.

El ADAS es un test que evalúa rendimiento y que consta de 21 ítems segregados en dos subescalas: cognitiva (ADAS-Cog) y conductual (ADAS-Noncog), de 11 y 10 ítems respectivamente. En la práctica se ha hecho muy popular el uso del ADAS-Cog, mientras que el ADAS-Noncog ha reemplazado por escalas conductuales más generales como Inventario Neuropsiquiátrico o NPI-Q (NeuroPsiquiatric inventary questionary). Éste, es un instrumento de aplicación relativamente breve que mide una serie de síntomas habituales en función de la frecuencia de aparición y la intensidad con que aparecen, aportando también una puntuación global que es la suma de las anteriores.

1.2. Técnicas Tomograficas empleadas en medicina nuclear

El desarrollo de la tomografía computerizada (CT: computed tomography) a principios de los años setenta revolucionó el campo de la radiología médica. Por primera vez se obtuvieron imágenes tomográficas de las estructuras internas del cuerpo humano. El primer sistema CT práctico fue desarrollado en 1971 por Dr. G.N. Hounsfield en Inglaterra con objeto de obtener una imagen del cerebro. Las proyecciones se tomaron en aproximadamente 5 minutos, y la imagen tomográfica se reconstruyó en 20 minutos. Desde entonces, la tecnología CT ha evolucionado notablemente y se ha convertido en un procedimiento estandarizado para diagnóstico mediante imágenes de casi todos los órganos del cuerpo humano en miles de hospitales y centros médicos de todo el mundo. En la actualidad los datos de las proyecciones se toman en aproximadamente 1 segundo, y la imagen se reconstruye en 3-5 segundos.

La medicina nuclear se puede definir como la práctica de hacer a los pacientes radiactivos con propósitos de diagnóstico o terapéuticos. La radiactividad se inyecta vía intravenosa o es ingerida. La circulación de una sustancia radiactiva en el cuerpo humano es la característica fundamental que distingue a la medicina nuclear de la radiología o la radiación oncológica en la mayoría de sus formas. A continuación se introducen las dos técnicas tomográficas más empleadas en medicina nuclear y que servirán de base para presentar las motivaciones de este proyecto.

La tomografía computerizada por emisión de un solo fotón (SPECT) y la tomografía por emisión de positrones (PET) son dos modalidades de imágenes no invasoras que proporcionan una distribución tridimensional de un radiofármaco en el cuerpo humano. La información clínica que estas imágenes proporcionan se relaciona con los procesos bioquímicos y fisiológicos en los que los radiofármacos se encuentran involucrados. La información funcional o metabólica aportada es lo que diferencia a esta modalidad de otras que únicamente proporcionan información anatómica o estructural. Esta ventaja se ha aprovechado en el diagnóstico temprano y en el tratamiento de enfermedades basadas en los cambios metabólicos o fisiológicos que puedan aparecer con anterioridad a que sean detectables modificaciones anatómicas o estructurales de tejidos internos u órganos. El elevado número de aplicaciones clínicas de este tipo de imágenes de emisión se debe a tres avances tecnológicos fundamentales: 1) La aparición de radiofármacos con átomos radioactivos que pueden ser administrados a los seres humanos de forma segura, 2) los avances técnicos alcanzados en el diseño de detectores de radiación, y 3) el aumento en las prestaciones de los computadores modernos y el desarrollo de técnicas efectivas para el procesado de imágenes y modelado biocinético.

1.2.1. SPECT

La tomografía basada en la emisión de un solo fotón (del inglés "single photon emission computed tomography" SPECT) es una técnica ampliamente usada para el estudio de las propiedades funcionales del cerebro (English and Childs, 1996; Ayache, 1996). Representa una modalidad de imagen médica que combina los principios de la medicina nuclear con las técnicas CT. Si la tomografía computerizada emplea rayos X, SPECT utiliza productos farmacéuticos radioactivos (por ejemplo el Tc-99m ethyl cys-

teinate dimer (ECD)) que se distribuyen en los diferentes tejidos internos u órganos del cuerpo humano. La distribución de los radiofármacos depende de sus propiedades biocinéticas y del estado anormal del paciente. Los fotones gamma emitidos por la fuente radioactiva son detectados mediante detectores de radiación similares a los utilizados en medicina nuclear. Los métodos de reconstrucción necesitan diferentes proyecciones tomadas desde diferentes vistas del paciente. Estos datos permiten generar imágenes que muestran la distribución del radiofármaco y son de gran utilidad en el diagnóstico de una gran variedad de enfermedades. Cuando se comparan con imágenes obtenidas mediante técnicas convencionales, las imágenes SPECT proporcionan un mejor contraste en la distribución de los radiofármacos en tejidos y órganos en medicina nuclear.

Una vez reconstruida a partir de las cuentas que proporciona la gammacámara (véase figura 1.1) y una correcta normalización de la imagen de SPECT, tomada por ejemplo mediante Tc-99m ethyl cysteinate dimer (ECD) como trazador, se obtiene un mapa de activación que representa la intensidad local del flujo sanguíneo regional (rCBF). Por lo tanto, esta técnica es aplicable al diagnóstico de enfermedades neurológicas como por ejemplo la enfermedad de Alzheimer (AD). (Hellman et˜al., 1989; Holman et˜al., 1992; Johnson et˜al., 1993; Stoeckel et˜al., 2001, 2004; Fung and Stoeckel, 2007). Es necesario, por tanto, el desarrollo de herramientas para mejorar la precisión en el diagnóstico en la etapa inicial de la enfermedad, cuando el paciente puede beneficiarse de los nuevos fármacos disponibles.

1.2.2. PET

La historia de la tomografía por emisión de positrones (PET: "Positron-Emission Tomography") se remonta a los años 50, cuando investigadores estadounidenses en Boston demostraron las posibilidades de una clase particular de sustancias radiactivas para producir imágenes médicas. Se demostró entonces que los fotones de alta energía producidos por aniquilación de positrones podrían emplearse para describir, en tres dimensiones, la distribución fisiológica de compuestos químicos. A mediados de los años ochenta, la técnica PET se había convertido ya en una valiosa herramienta de diagnóstico en medicina y para el estudio dinámico del metabolismo humano. En la actualidad, debido a la mayor sensibilidad de PET frente a las imágenes

Figura 1.1: Gamma Cámara Picker Prism 3000

obtenidas por resonancia magnética (MRI), se emplea en el estudio de los neuro-receptores en el cerebro y otros tejidos del cuerpo humano.

El proceso de obtención de imágenes PET se inicia con la ingesta de un trazador de la actividad metabólica; una molécula que contiene un isótopo emisor de positrones (por ejemplo, ^{11}C, ^{13}N, ^{15}O ó ^{18}F). Pasados unos minutos, el isótopo se acumula en áreas con las que la molécula tiene una cierta afinidad. Por ejemplo, la glucosa etiquetada con ^{11}C se acumula en cerebro o en tumores, donde la glucosa se emplea como fuente principal de energía.

Los scans PET que usaremos en este trabajo son un ejemplo de modalidad tridimensional no invasiva de imagen funcional cerebral, que mide la tasa de metabolismo de glucosa en el cerebro a través del trazador [^{18}F] Fluorodeoxyglucosa. En la enfermedad del Alzheimer, regiones características muestran un decrecimiento en el metabolismo de glucosa, especificamente regiones bilaterales en los lóbulos temporal y parital, cíngulo posterior y precunei y también en el cortex frotal y el conjunto global del cerebro en casos de afección severa. (de˜Leon et˜al., 1983; Foster et˜al., 1983, 1984; Chase et˜al.,

1984; Duara et˜al., 1986; McGeer et˜al., 1990; Minoshima et˜al., 1994, 1995; Ibañez et˜al., 1998; Hoffman et˜al., 2000; Kogure et˜al., 2000; Alexander et˜al., 2002; Mosconi et˜al., 2008; Langbaum et˜al., 2009). La severidad de la demencia ha sido correlacionada con tasas bajas de metabolismo de glucosa, tanto estableciendo correlación entre medidas de capacidad de las funciones cognitivas, como por ejemplo el Mini-Mental State Exam (MMSE), y la reducción de la tasa de metabolismo cerebral de glucosa (cerebral metabolic rate for glucose (CMRgl)) entre todos los sujetos del estudio, o comparando subgrupos basándose en medidas de la severidad de la enfermedad como MMSE, Clinical Dementia Rating (CDR), o puntuación en la Global Deterioration Scale (GSD) (Chase et˜al., 1984; Minoshima et˜al., 1995; Foster et˜al., 1984; Silverman et˜al., 2001).

CAPÍTULO 2

Diagnóstico Asistido por Computador

Hasta la fecha se han propuesto numerosos sistemas de ayuda al diagnóstico de enfermedades neurológicas, con el objetivo de analizar imágenes de tipo SPECT u otros tipos de imágenes funcionales. La aproximación univariada más relevante esta basada en SPM (del inglés "Statistical Parametric Mapping") y sus numerosas variantes (Friston et˜al., 2007). Grosso modo, SPM consiste en hacer un test estadístico univariado a nivel de vóxel (unidad de volumen mínimo de la imagen), por ejemplo un test "t"-student de dos muestras, que compara los valores del vóxel de la imagen bajo estudio con un grupo de pacientes normales que representan la muestra "control". A continuación, los vóxeles relevantes de este test son inferidos usando la teoría de campos aleatorios (Adler, 1981). Su marco de actuación fue pensado para el análisis de estudios de imágenes SPECT y PET, pero actualmente se aplica principalmente para el análisis de la MRI (del inglés "Magnetic Resonance Imaging") funcional.

Sin embargo, SPM no está estrictamente diseñada para resolver el problema del diagnóstico automático usando exclusivamente un paciente de estudio sino para la comparación de conjuntos de imágenes a las cuales se les asigna una etiqueta implícitamente (como es el caso diagnóstico que nos ocupa). De hecho, su aplicación en este contexto proporciona resultados de clasificación pobres (semejantes a nuestra metodología de referencia (Stoeckel et˜al., 2001)) dado que una de las poblaciones consiste en un único individuo (estimación sesgada de la media de la población), y la otra consiste en un conjunto de individuos normales (el test "t" no incluye ninguna información sobre la patología bajo estudio) (Stoeckel et˜al., 2001). Más aún, este método sufre los inconvenientes de las aproximaciones locales y univariadas.

Por otro lado, se ha propuesto otras aproximaciones multivariadas como ManCova, que consideran como una observación todos los vóxeles de un solo "scan" con el objetivos de hacer inferencias sobre los efectos de activación distribuidos. La importancia de ellas radica en que los efectos debidos a activaciones, los efectos indefinibles ("confounding effects") y los efectos de error son evaluados estadísticamente en términos de efectos a nivel de vóxel y también a nivel de las interacciones entre vóxeles (Frackowiak et˜al., 2003). No obstante, con estas técnicas uno no puede hacer inferencias estadísticas sobre cambios específicos regionales, y, aun más importante, requieren un número de observaciones (scans) que sea mayor que el número de componentes de la observación multivariada (vóxeles). Obviamente esta no es la situación en la que nos encontramos cuando trabajamos con estudios de imágenes funcionales (SPECT, PET, fMRI).

El trabajo del grupo se centra en el contexto de las aproximaciones *supervisadas* y multivariadas donde se plantea un nuevo método cuantitativo para evaluación de imágenes funcionales. En este campo, en el que el grupo BIP es referencia, la clasificación se realiza habitualmente mediante la definición de vectores de características que representan los rasgos más relevantes de las diferentes imágenes, por ejemplo las de SPECT y mediante el entrenamiento de un clasificador dado un conjunto de muestras conocidas (Illán et˜al., 2009; Górriz et˜al., 2008), etc. Después del proceso de entrenamiento el clasificador (que incluye la capacidad de generalización del sistema) se aplica a nuevos casos de test para distinguir entre controles sanos y enfermos. En esta régimen de entrenamiento y test se asume como hipótesis de partida que las etiquetas de entrenamiento y test son válidas por lo que una precisión ele-

vada en la clasificación es equivalente a un diagnóstico efectivo del paciente bajo estudio.

El conjunto de clasificadores usados por los sistemas CAD están basados en distintas funciones analíticas (por ejemplo en su complejidad) que se ajustan mediante datos de entrenamiento en base a distintos procedimientos. De entre esos procedimientos destacamos, por su robustez, el uso de máquinas de vectores soporte (SVM del inglés "support vector machines"). Desde su introducción en los setenta, las SVMs han marcado el comienzo de una nueva era en el paradigma del aprendizaje a partir de muestras (Vapnik, 1998). SVMs han atraído recientemente la atención de la comunidad científica en el campo de reconocimiento de patrones dado el alto número de avances a nivel teórico y computacional derivados de la Teoría del Aprendizaje Estadístico (TAE) (Vapnik, 1998) desarrollada por Vladimir Vapnik en los laboratorios de AT&T. Estas técnicas han sido usadas de manera satisfactoria en un gran número de aplicaciones entre las que destacamos la detección de actividad de voz (VAD) (Ramírez et˜al., 2006b), recuperación de imágenes basadas en contenido (Tao et˜al., 2006), clasificación de texturas (Kim et˜al., 2002) o el diagnóstico basado en imagen (Illán et˜al., 2009; Fung and Stoeckel, 2007).

La ventaja de este tipo de aproximación para el diagnóstico clínico basada en la TAE es que no necesario ningún tipode conocimiento a priori acerca de la enfermedad bajo estudio y que el método que es automático, es aplicable a cualquier otro tipo de patología neurológica o técnica de imagen cerebral. En al aproximación básica, que nos sirve de referencia, las intensidades de los vóxeles I_n de la imagen funcional, por ejemplo SPECT, son directamente usados para construir los vectores de características $v = (I_1, \ldots, I_N)$, véase por ejemplo (Stoeckel et˜al., 2001, 2004). Esta aproximación se ha demostrado en las anteriores referencias ser igual de exacta que el diagnóstico visual usando como apoyo SPM, sin embargo presenta los mismo problemas que las aproximaciones multivariadas como Mancova: incluso después de bajar la resolución de la imagen sub-muestreando, y tras aplicar una *máscara* cerebral seguimos teniendo un problema en la clasificación de dimension $N \sim 10000$ en los vectores de características. Por lo tanto, el tamaño de estos vectores es muy superior al número de muestras (50-100 es un número realista), lo cual conduce al llamado problema de tamaño muestral pequeño (del inglés "small sample size problem") (Duin, 2000).

CAPÍTULO 3

Mapa Paramétrico Estadístico (SPM)

En el ámbito de la investigación médica, el análisis de neuroimágenes funcionales (PET, SPECT y fMRI) mediante técnicas de cuantificación estadística permite el estudio de diversos procesos cerebrales, patológicos o cognitivos. Una herramienta de creciente uso para este fin el conocido programa SPM (del inglés "Statistical Parametric Mapping") gracias a su amplia disponibilidad y el gran abanico de estudios estadísticos que permite realizar. Sin embargo, el desconocimiento de los fundamentos teóricos en los que se basa puede conducir fácilmente a resultados imprecisos e incluso a conclusiones erróneas. Esta sección presenta brevemente dichos principios teóricos y discute los principales puntos críticos en la utilización del método sin detallar los fundamentos matemáticos en los que está basada la citada herramienta. En este sentido la finalidad de SPM es la realización de mapas de estadísticos paramétricos para la búsqueda de efectos de interés presentes en imágenes funcionales PET, SPECT o fMRI. Desde su primera aparición en 1991, la comunidad de investigadores de neuroimagen funcional ha aceptado y utilizado ampliamente las actualizaciones de 1996 y 1999, gracias a que pro-

porcionan una gran flexibilidad en el diseño de los experimentos que pueden analizarse (Friston et˜al., 2007). SPM se utiliza actualmente en departamentos de psiquiatría, psicología, neurología, radiología, medicina nuclear, farmacología, ciencias cognitivas y del comportamiento, bioestadística y física biomédica de todo el mundo para la investigación de enfermedades mentales, cuantificación de efectos farmacológicos, estudios cognitivos, realización de análisis longitudinales, estudios intersujeto, e incluso morfométricos. En un estudio estadístico mediante SPM los puntos clave son la elección del método de normalización en intensidad, la normalización espacial, el sistema de coordenadas empleado y la interpretación de la significación estadística de los resultados. Estos serán las cuestiones principales a tratar en esta sección presentando las alternativas y soluciones posible a ellas.

3.1. Preprocesado en SPM

El preprocesado empleado en el paradigma basado en SPM es el empleado en todos los métodos de diagnóstico basados en imagen. Por ejemplo, éstos requieren que las imágenes usadas en el procedimiento sean comparables vóxel a vóxel estableciendo una correspondencia exacta entre posición anatómica y posición del vóxel en la imagen. Este preprocesado consta de tres etapas: *realineado*, *normalización espacial* y *filtrado espacial*.

En la etapa de *realineado* se consigue corregir la diferente colocación de la cabeza en distintas imágenes de un mismo sujeto dentro del dispositivo de imagen (PET, SPECT, fMRI). Para corregirla, se aplican las traslaciones y rotaciones adecuadas que compensen esta diferencia, de modo que las imágenes coincidan en el mismo espacio común. en el caso de diagnóstico que nos concierne no se aplica tal transformación dado que solo se dispone de una imagen de cada paciente. Estas transformaciones sencillas parten del hecho de que el cerebro a realinear tiene la misma morfología en cada adquisición. Sin embargo cuando disponemos de distintos pacientes ese no es el caso para ello aplicamos la etapa de *normalización espacial*. En efecto, para realizar un análisis vóxel a vóxel, los datos de distintos sujetos deben corresponderse con un espacio anatómico estándar que permite la comparación entre sujetos y la presentación de los resultados de un modo convencional.

En esta etapa se realiza una deformación elástica (Salas-González et˜al., 2008; Salas-Gonzalez et˜al., 2008; Friston et˜al., 2007) de las imágenes de modo que concuerden con un patrón anatómico estandarizado. Para que la transformación espacial sea correcta, las imágenes deben ser razonablemente similares al patrón utilizado, tanto morfológicamente como en contraste. Este patrón se obtiene promediando un conjunto de controles normales de manera que se obtiene una plantilla o "template" suave que sirve como referencia. De este modo, se ponen en correspondencia cada una de las regiones cerebrales de cada sujeto con una localización homóloga en el espacio estándar. Esta normalización, además de permitir la comparación vóxel a vóxel de las imágenes, también facilita la localización de las áreas funcionales. El concepto de sistematizar la localización cerebral de las regiones funcionales se debe originalmente a Talairach (J. and P., 1988), y si bien SPM presenta los resultados finales mediante este método, el sistema de coordenadas empleado para informar acerca de las localizaciones no es el mismo que el que aparece en el atlas de Talairach, lo que puede inducir a error. El filtrado es un proceso por el cual los vóxeles se promedian con sus vecinos, produciendo un suavizado de las imágenes, más o menos pronunciado en función de un parámetro denominado FWHM (del inglés "Full Width at Half Maximum", Amplitud Total a Media Altura). La FWHM tiene unidades espaciales y mide el grado de suavizado: a mayor FWHM, mayor suavizado. El suavizado de las imágenes tiene diversos objetivos. En primer lugar, aumenta la relación señal/ruido, ya que elimina fundamentalmente las componentes ruidosas de la imagen. Otro motivo que hace conveniente suavizar las imágenes es que así se garantiza que los cambios entre sujetos se presentarán en escalas suficientemente grandes como para ser anatómicamente significativas, una vez efectuado una normalización en intensidad.

3.2. Análisis Estadístico

Mediante SPM es posible realizar numerosos tests estadísticos, como regresiones, tests t de Student, tests F y análisis de varianza (Anova) incluyendo covariables y permitiendo el modelado de interacciones entre ellas (Friston et˜al., 2007). Todos estos tipos de análisis pueden ser englobados en un modelo general (el modelo lineal general o GLM), que es el utilizado por SPM

para efectuar los cálculos matemáticos. La formulación del GLM se basa en dos conceptos: la matriz de diseño y los contrastes.

Los estudios estadísticos que pueden efectuarse mediante SPM pueden ser divididos fundamentalmente en dos tipos: estudios paramétricos o factoriales y estudios categóricos o sustractivos. Los primeros estudian la relación entre las imágenes PET y un parámetro, como puede ser la edad, una escala de síntomas o el resultado de un test cognitivo. Los estudios categóricos se utilizan para poner de manifiesto diferencias entre grupos, definidas por variables categóricas. Mediante el uso del GLM somos capaces de introducir de manera matemática el estudio estadístico en cuestión, que en el caso del diagnóstico de un paciente a partir de una sola imagen de test, es la definición de la matriz de diseño de la siguiente forma. Dentro de los estudio categóricos es necesaria una columna para determinar la pertenencia a cada uno de los grupos (categoría normal frente a Alzheimer). Por ejemplo, si se desea comparar un grupo de pacientes con otro de control, las filas correspondientes a los pacientes en la matriz de diseño, tendrán un uno en la columna de pacientes y un cero en la de controles, y viceversa.

Una vez establecido el modelo (véase figura 3.1) , SPM ya puede estimar de forma automática la contribución de cada efecto de forma separada. Esto permite diferenciar entre efectos de "interés" (como el efecto de grupo) y efectos "correctores" (por ejemplo efecto de la edad en estudios paramétricos), así como por diferencias entre las medias de dos factores. Mediante el GLM, esto se realiza mediante la definición de un "contraste" que se define como un vector. La longitud de este vector es igual al número de efectos incluidos en la matriz de diseño, de modo que cada efecto se pondera por su elemento correspondiente. Si el efecto es corrector, entonces se pondera con un cero en el vector contraste. En caso de que el efecto sea paramétrico, el contraste determina si la correlación buscada es positiva, mediante un "1", o bien negativa mediante un "-1", en la posición correspondiente a ese efecto en el vector contraste. En caso de efectos categóricos los contrastes deben cumplir una condición importante: la suma de todos los pesos en el contraste en las columnas de efectos categóricos debe ser igual a cero. En el ejemplo, la pertenencia al grupo de pacientes contribuye con peso negativo (menor metabolismo que controles), el grupo de controles contribuye con peso positivo (mayor metabolismo que pacientes). De este modo, el vector de contraste sería [1 -1], ya que la suma de los efectos 1 y 2 debe ser cero. En el caso

contrario, para comprobar qué regiones presentan un metabolismo mayor en pacientes, el contraste sería [-1 1].

Finalmente, SPM realiza el test estadístico (un test t o un test F), descrito mediante la matriz de diseño y el contraste, en todos los vóxeles de la imagen de forma independiente (véase figura 3.1). El resultado es una imagen cuyo valor en cada vóxel es el resultado del test estadístico y a la que se denomina mapa paramétrico estadístico

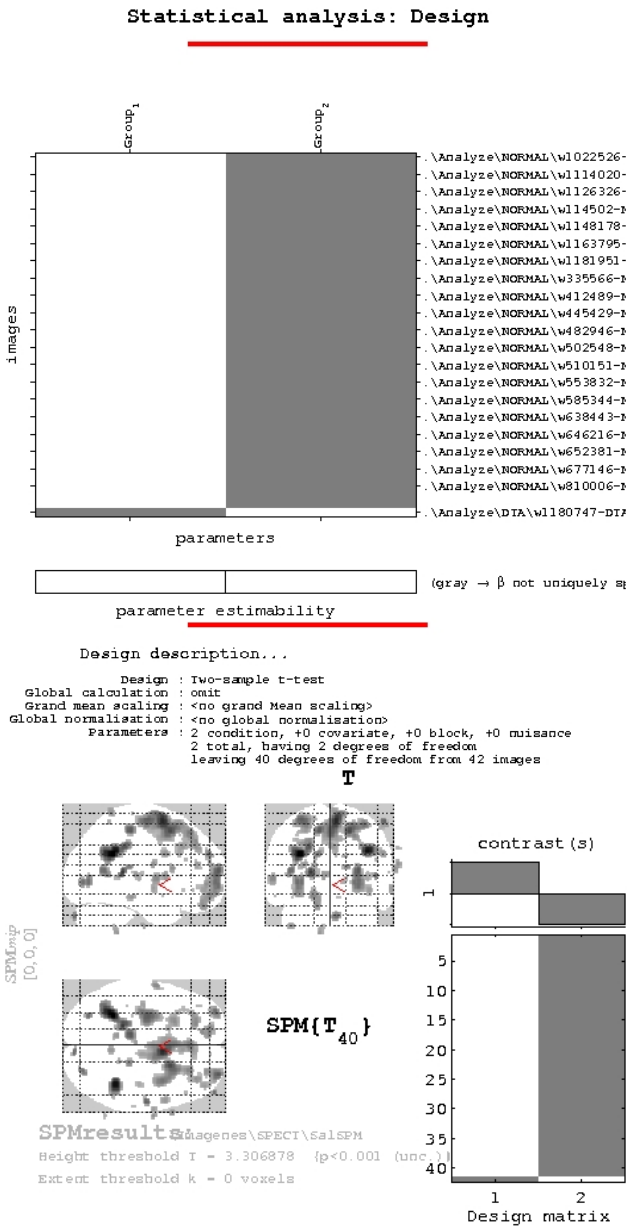

Figura 3.1: Resultados de SPM para clasificación con el modelo descrito en el texto

CAPÍTULO 4

Aprendizaje Estadístico Supervisado en Neurociencias

Una rama importante de los métodos de ayuda al diagnóstico médico por ordenador se basa en las técnicas conocidas como *métodos de clasificación*, en el contexto de la Teoría de Aprendizaje Estadístico. Estas técnicas consisten en la extracción de cierta información a partir de un conjunto de datos (aprendizaje automático), que es empleada posteriormente para hacer diagnósticos de nuevos pacientes. En el sentido más amplio, cualquier método que incorpora información de un conjunto de muestras para el diseño de una función clasificadora emplea aprendizaje. Con aprendizaje nos referimos, en este contexto, a un procedimiento algorítmico para reducir el error de la clasificación en el conjunto de entrenamiento. Dado que casi todos los problemas de reconocimiento de patrones son bastante complejos se suele invertir la mayor parte del tiempo en la fase de aprendizaje.

Una de las formas más populares de aprendizaje automático se denomina con el sufijo de supervisado. El *aprendizaje supervisado* implica el uso de datos de entrenamiento sobre los que se conoce la clase a la que pertenecen. En el caso de diagnóstico, las clases vendrán determinadas mediante una etiqueta binaria (± 1), que determinará si el paciente padece o no la enfermedad. El proceso de etiquetado dependerá de la base de datos en concreto, aunque en general se asumirá que el error en el etiquetado es despreciable. Así, las imagenes cerebrales contendrán la información de diferentes procesos que tienen lugar en el cerebro en forma de distribuciones de intensidad 3 dimensionales y sus etiquetas correspondientes. La distribución de intensidad I corresponderá a diferentes aspectos de la actividad funcional del cerebro, ya sea una imagen SPECT o PET (como por ejemplo la tasa de consumo de glucosa), y estará discretizada a un número M finito de elementos de volumen llamados *voxels*. Podremos agrupar el conjunto de datos originales a partir de los valores que toma la intensidad en los diferentes puntos del cerebro, para formar vectores:

$$\mathbf{x} = (I_1, I_2, ..., I_M) \tag{4.1}$$

que nos permitirán construir los objetos del aprendizaje supervisado: los vectores de características.

Definición 1: El conjunto de datos experimentales o muestras en el proceso de clasificación estará formado por un conjunto de *vectores de características* $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, ..., n$, siendo m la dimensión del *espacio de características* $\mathcal{H}$. Este conjunto de vectores de características se dividirá en 2 subconjuntos: datos de entrenamiento y datos de test:

- Diremos que un vector de características $\mathbf{x}_i$ es un *vector de entrenamiento* si pertenece al subconjunto de entrenamiento $\mathcal{X} \subset \mathbb{R}^m$. Expresado matemáticamente, $\mathbf{x}_i \in \mathcal{X}$.
- Diremos que un vector de características $\bar{\mathbf{x}}_j$ es un *vector de test* si pertenece al subconjunto de test $\mathcal{Y} \subset \mathbb{R}^m$. Expresado matemáticamente, $\bar{\mathbf{x}}_j \in \mathcal{Y}$.

- La unión de estos dos subconjuntos formará el espacio de características $\mathcal{H}$, es decir, $\mathcal{H} = \mathcal{X} \cup \mathcal{Y}$. En nuestro caso concreto, asumiremos que $\mathcal{H}$ es sencillamente $\mathbb{R}^m$.

Los vectores de características se obtendrán del conjunto de datos originales (4.1), tomandose toda o parte de la información contenida en ellos. A menudo se obtendrá de ellos aquellas características más representativas para efectuar la clasificación, lo que supondrá una disminución del tamaño de los datos originales, cuestión que abordaremos en profundidad en la sección 4.3. Una vez obtenidos los vectores de características, será posible definir el clasificador:

Definición 2: Se define un clasificador como una función $f = f(\mathbf{x}_i, \omega)$, dependiente de unos parámetros ω y construida a partir de los datos de entrenamiento n-dimensionales $\mathbf{x}_i \in \mathcal{X}$ y sus correspondientes etiquetas y_i, tal que clasifica un nuevo vector de test $\bar{\mathbf{x}}_j \in \mathcal{Y}$, asignándole un valor $z_j \in \{\pm 1\}$ correpondiente a cada clase:

$$\begin{array}{rccl} f: & \mathbb{R}^n & \longrightarrow & \{\pm 1\} \\ & \mathbf{x}_j \in \mathcal{Y} & \longmapsto & f(\mathbf{x}_j, \omega) = z_j \end{array} \tag{4.2}$$

El objetivo del proceso de aprendizaje supervisado es construir una función o clasificador de manera que f categorice correctamente nuevos patrones $\mathbf{x}_j$. Así, el proceso de clasificación se dividir en dos etapas: el entrenamiento y el test. En la etapa de entrenamiento se define el clasificador de acuerdo con un conjunto de vectores de características de entrenamiento $\mathbf{x}_i \in \mathcal{X}$, cuyas etiquetas son conocidas. La forma exacta del clasificador dependerá del modelo propuesto para la tarea de clasificación, y los parmámetros desconocidos del modelo serán estimados empleando los patrones de entrenamiento $\mathbf{x}_i \in \mathcal{X}$, $i = 1, ..., l$. El algoritmo de aprendizaje automático consistirá en estimar los parámetros ω, de manera que el error de clasificación sobre el conjunto de entrenamiento sea mínimo. Una vez definido, el clasificador se emplea para establecer categorías sobre muestras desconocidas pertenecientes al conjunto de test.

En la sección 4.2 estudiaremos varios modelos de clasificador clásicos, aunque nuestro interés se centrará en el clasificador basado en la metodología SVM, y que estudiaremos en profundidad en el capítulo 5.

La elección de los conjuntos de entrenamiento y test es un problema ampliamente estudiado en la literatura. Una condición necesaria es que los conjuntos sean independientes, es decir, que no existan elementos del conjunto de test que hayan sido utilizados en el entrenamiento del clasificador. Expresado matemáticamente requiere que:

$$\mathcal{X} \bigcap \mathcal{Y} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \bigcap \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_n\} = \oslash \tag{4.3}$$

Además, es importante que estos conjuntos sean muestras representativas de la población que se quiere estudiar y clasificar, ya que esto permitiría construir el clasificador adecuadamente y dar una buena respresentación de su comportamiento.

El método que permite evaluar los resultados de cualquier análisis estadístico es conocido como Validación, y se estudiará en la sección 4.5 como método para evaluar el comportamiento de un clasificador. Uno de los aspectos interesantes a evaluar sobre la clasificación es cómo los resultados pueden ser generalizados a un conjunto de datos independientes. La *generalización* es un concepto central en el diseño de clasificadores y hace referencia a la capacidad de un clasificador construido a partir de un conjunto muestral de entrenamiento determinado, para describir la esctructura subyaciente de la población total, y no de la muestra concreta. Así, un clasificador con buena capacidad de generalización será capaz de operar correctamente con nuevos datos independientes. Otros aspectos a valorar sobre el rendimiento de un clasificador se estudiarán en la sección 4.4.

Es posible que el conjunto de vectores de características quede representado de una manera más simple en un espacio diferente al original. En estos casos, la eficiencia del clasificador mejorará cuando se realice la transformación oportuna, a través de las funciones conocidas como *kernels*. El criterio a seguir lo marcará el *Teorema de Mercer*, estudiado en la sección 4.1, imponiendo las condiciones necesarias para que esta transformación pueda tener lugar.

Los métodos de clasificación, por tanto, son técnicas complejas en las que intervienen varios factores: definción del clasificador, extracción de características, aprendizaje o entrenamiento, y test. Uno de los problemas que afectará a algunos de estos factores, en el caso de la clasificación de imágenes médicas para la ayuda al diagnóstico, a la hora de construir un clasificador eficaz es el problema de el pequeño número de muestras, que se da debido al reducido número de pacientes que intervienen en los estudios en comparación con la gran cantidad de información contenida en las imagenes cerebrales. Estudiaremos en la sección 4.3 como abordarlo.

4.1. Métodos Kernel

Se puede pensar que, hablando en términos generales, un clasificador 'aprende' a encontrar similitudes entre los datos, incorporando en cierta manera el concepto de *similaridad*. De alguna manera, al construir un clasificador se comparan los datos $\mathbf{x}$ cuyas etiquetas son conocidas, con nuevos patrones $\mathbf{x}'$, y se categoriza el nuevo patrón en función de su similitud con los datos conocidos. En el caso de datos etiquetados binariamente la caracterización de esta similaridad es sencilla, ya que solo pueden darse dos situaciones: que las etiquetas sean iguales o diferentes. Sin embargo, la medida de esta similaridad es un asunto profundo que radica en el núcleo del campo del aprendizaje automático. Matemáticamente, podemos considerar que una medida de la similaridad viene dada por una función:

$$\begin{array}{rccc} k: & \mathcal{H}\times\mathcal{H} & \longrightarrow & \mathbb{R} \\ & (\mathbf{x},\mathbf{x}') & \longmapsto & k(\mathbf{x},\mathbf{x}') \end{array} \tag{4.4}$$

es decir, una función que, dadas dos muestras del espacio de características $\mathbf{x}$ y $\mathbf{x}'$, devuelve un número real que caracteriza su similaridad. Es natural asumir que no existe diferencia al comparar la similaridad entre $\mathbf{x}$ y $\mathbf{x}'$ que entre $\mathbf{x}'$ y $\mathbf{x}$, por lo que asumiremos que esta función es simetrica, es decir, $k(\mathbf{x},\mathbf{x}') = k(\mathbf{x}',\mathbf{x})$, para todo $\mathbf{x},\mathbf{x}' \in \mathcal{H}$.

Un ejemplo sencillo de una función de este tipo es el producto escalar en un espacio vectorial, que lleva asociado una noción de distancia. La medida de la similaridad en un espacio con un producto interno definido se traduce a

una medida de distancia, haciendo que se pueda establecer si dos objetos son similares a través de su cercanía[1]. Así, si $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$, se define el producto escalar o producto interno como:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}' = \sum_{i=1}^{m} x_i x_i' \tag{4.5}$$

con $\mathbf{x} = (x_1, ..., x_m)$ y $\mathbf{x}' = (x_1', ..., x_m')$. Gracias a la propiedad del producto interno $(\mathbf{x} \cdot \mathbf{x}) \geq 0$, se puede definir el concepto de norma de un vector como:

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x} \cdot \mathbf{x})} \tag{4.6}$$

lo que permite nutrir el espacio de características con los conceptos habituales del espacio Euclídeo de longitud, distancia y ángulo.

Estas ideas se pueden extrapolar a espacios más generales que $\mathbb{R}^m$, en los que también es posible definir un producto interno. Estos espacios, denominados espacios de Hilbert, son ánalogos al espacio Euclídeo en tanto que se puede definir en ellos los conceptos de longitud, distancia y ángulo, pero poseen características más generales. Son la generalización del espacio Euclídeo, reproduciendo el algebra vectorial de manera abstracta e incluyendo una noción de completitud.

Es posible que la noción de similaridad sea definida más apropiadamente en otro espacio V diferente al espacio de características original $\mathcal{H}$. Supondremos que este espacio V es un espacio de Hilbert, con su correspondiente producto interno. Primero, necesitaremos una representación vectorial de los vectores de características en este espacio, que se construirá a través del mapeo:

$$\begin{array}{rccc} \varphi: & \mathcal{H} & \longrightarrow & V \\ & \mathbf{x} & \longmapsto & \mathbf{v} = \varphi(\mathbf{x}) \end{array} \tag{4.7}$$

[1]En la sección 4.2.1 se introduce un clasificador que hace uso de estas ideas para categorizar los datos de entrada

Este mapeo definirá un nuevo espacio de características V y sus correspondientes vectores de características $\mathbf{v} \in V$, de manera que el producto escalar quedará mapeado a:

$$k(\mathbf{v}, \mathbf{v}') = \mathbf{v} \cdot \mathbf{v}' = \varphi(\mathbf{x}) \cdot \varphi(\mathbf{x}) = k(\varphi(\mathbf{x}), \varphi(\mathbf{x})) \tag{4.8}$$

de manera que un kernel es sencillamente una función que puede ser representada como un producto interno en algún espacio de Hilbert.

Esta tranformación permitirá seguir trabajando con las nociones habituales de distancia, longitud y ángulo, y las herramientas sencillas empleadas para trabajar con ellas (algebra lineal, geometría analítica,...) pero abriendo la posibilidad de utilizar un amplio rango de diferentes medidas de la similaridad, gracias a la libertad a la hora de elegir el mapeo φ. Uno puede preguntarse sobre las condiciones generales para la existencia de tal mapeo, cuya respuesta viene dado por el Teorema de Mercer:

Teorema. *Teorema de Mercer. Sea* $\mathbf{x} \in \mathbb{R}^m$ *y* φ *una función de mapeo*

$$\begin{array}{rccl} \varphi: & \mathbb{R}^m & \longrightarrow & V \\ & \mathbf{x} & \longmapsto & \mathbf{v} = \varphi(\mathbf{x}) \end{array} \tag{4.9}$$

donde V *es un espacio de Hilbert. Entonces, la operación del producto interno tiene una representación equivalente*

$$\sum_i \varphi_i(\mathbf{x})\varphi_i(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$$

donde $\varphi_i(\mathbf{x})$ *es la componente* i *del mapeo* $\varphi(\mathbf{x})$ *de* $\mathbf{x}$, *y* $k(\mathbf{x}, \mathbf{x}')$ *es una función simétrica que satisface la siguiente condición:*

$$\int k(\mathbf{x}, \mathbf{x}')g(\mathbf{x})g(\mathbf{x}')d\mu(\mathbf{x})d\mu(\mathbf{x}') \geq 0 \tag{4.10}$$

para cualquier función de cuadrado integrable $g \in L_2$, es decir, tal que

$$\int g(\boldsymbol{x})^2 d\mu(\mathbf{x}') < +\infty \tag{4.11}$$

donde $d\mu(\boldsymbol{x}')$ *es la medida de Borel-Lebesgue.*

La afirmación contraria del teorema es también cierta, es decir, para cualquier $k(\mathbf{x}, \mathbf{x}')$ que satisfaga (4.10) y (4.11) existe un espacio en el cual k defina un producto interno. Puesto que nosotros partimos de una representación del conjunto de datos (4.1) en un espacio vectorial con producto interno, se puede entender el mapeo (4.9) como un isomorfismo entre espacios de Hilbert, por lo tanto es simpre posible definir un mapeo lineal uno-a-uno entre espacios que preserve la estructura de producto interno. En estos casos, la definición de la función kernel se conoce como *kernel trick*, en la que en lugar de construir una función kernel, se establece una transformación entre funciones kernel. Más aun, estaremos interesados por funciones g diferenciables, por lo que la medida de Borel-Lebesgue de (4.10) y (4.11) se puede entender como la medida normal de la integral de Riemann. Sin embargo, existen construcciones mucho más generales que permiten usar estas técnicas para casos en los que el espacio de características original no sea un espacio vectorial (Vapnik, 1998).

Gracias al teorema de Merecer podemos estar seguros de se puede establecer un mapeo a un nuevo espacio que permita definir una noción de similaridad. Normalmente este nuevo espacio implicará un aumento de la dimensión del espacio de características, para proporcionar un marco en el que las clases se separen más facilmente (ver (5.2). Lo que el teorema de Mercer no revela sin embargo es cómo encontrar este espacio. Es decir, no tenemos una herramienta general para construir el mapeo $\varphi(\cdot)$ una vez que conocemos el producto interno del correspondiente espacio.

En la práctica, ha resultado que los mapeos no lineales producen estructuras suficientemente interesantes para resolver problemas complejos, siendo algunos ejemplos típicos de kernels usados en aplicaciones de reconocimiento de patrones:

- Polinómicos:

$$k(\mathbf{x}, \mathbf{x}') = [\gamma(\mathbf{x} \cdot \mathbf{x}') + c]^d. \tag{4.12}$$

- Funciones de base radial (RBF):

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2). \tag{4.13}$$

- Tangente hiperbólica:

$$k(\mathbf{x}, \mathbf{x}') = \tanh(\gamma(\mathbf{x} \cdot \mathbf{x}') + c). \tag{4.14}$$

  Para valores apropiados de γ y c de modo que las condiciones de Mercer se cumplan.

4.2. Clasificadores estadísticos

A la hora de elegir el modelo sobre el cual construir el clasificador surgen diferentes cuestiones a tener en cuenta, como la robustez del algoritmo de aprendizaje frente a la variación de los parámetros del modelo, la convergencia del mismo en un tiempo razonable, su dependencia al número de patrones de entrenamiento, característica de entrada, su capacidad de generalización o de simplificación, favoreciendo soluciones sencillas frente a complejas.

Existe un amplio rango de algoritmos de descenso de gradiente para modificar los parámetros de un clasificador de manera que se reduzca cierta medida del error, dando lugar un nuevo campo denominado reconocimiento estadístico de patrones que se encarga de su estudio. Para nuestros propósitos empleamos un conjunto de clasificadores clásicos que se definen en base a una minimización de la función de coste error, aunque nos centraremos posteriormente en el estudio de los clasificadores de SVM.

Una elección apropiada de un clasificador debe tener en cuenta la distribución de las clases en el espacio de características. Por ejemplo, un clasificador lineal que defina un hiperplano en el espacio de características no será apropiado para separar dos clases representadas por dos círculos concéntricos. Sin embargo, la elección de un clasificador no lineal complejo no siempre garantiza obtener los mejores resultados. De hecho, cuanto mayor sea la complejidad de la función que define el clasificador, más parámetros habrán de ser estimados para definirlo, lo que puede ser problemático sobre todo si se cuenta con un número reducido de vectores de entrenamiento.

Presentaremos a continuación dos clasificadores sencillos que nos permitirán ilustrar muchas de las ideas introducidas hasta el momento. Ambos definirán un hiperplano en el espacio de características, lo que les encasilla en el grupo de clasificadores lineales. Los clasificadores lineales actuando sobre $\mathbf{x} \in \mathbb{R}^m$ se pueden escribir como:

$$f_{lineal}(\mathbf{x}) = \mathbf{x} \cdot \omega + \omega_0 \tag{4.15}$$

donde el vector ω y el escalar ω_0 forman un conjunto de $m + 1$ parámetos que han de ser estimados. A continuación se muestran dos métodos para estimarlos.

4.2.1. Nearest Mean

El *clasificador "nearest mean"* (media más cercana) usa los vectores de características medios μ_1 y μ_2 de las muestras de ambas clases $\mathcal{C}^1$ y $\mathcal{C}^2$, definidos como:

$$\mu_j = \frac{1}{N_j} \sum_{\{i|y_i \in \mathcal{C}^j\}} \mathbf{x}_i,\ j = 1, 2 \tag{4.16}$$

donde N_j es el número de vectores de características $\mathbf{x}_i$ cuya etiqueta y_j pertenece a la clase $\mathcal{C}^j$. En el caso de diagnósitico del Alzheimer, las clases serán sencillamente controles normales y pacientes AD.

El modelo del clasificador de media mas cercana asigna a un vector de características $\mathbf{x}'$ de una muestra desconocida la clase o etiqueta con vector media más cercano. El clasificador por lo tanto se define como:

$$f_{\mathrm{nm}}(\mathbf{x}') = (\mathbf{x}' - \mu_2)^T(\mathbf{x}' - \mu_2) - (\mathbf{x}' - \mu_1)^T(\mathbf{x}' - \mu_1), \tag{4.17}$$

donde la muestra se etiqueta como 'normal' si $f_{\mathrm{nm}}(\mathbf{x}') > 0$ y 'Alzheimer' si $f_{\mathrm{nm}}(\mathbf{x}') < 0$. Vemos como en (4.17) interviene el producto interno a través de producto de matrices $(\mathbf{x}' - \mu_i)^T(\mathbf{x}' - \mu_i)$, lo que supone un ejemplo de criterio de similiaridad definido a través de un concepto de distancia, en este caso

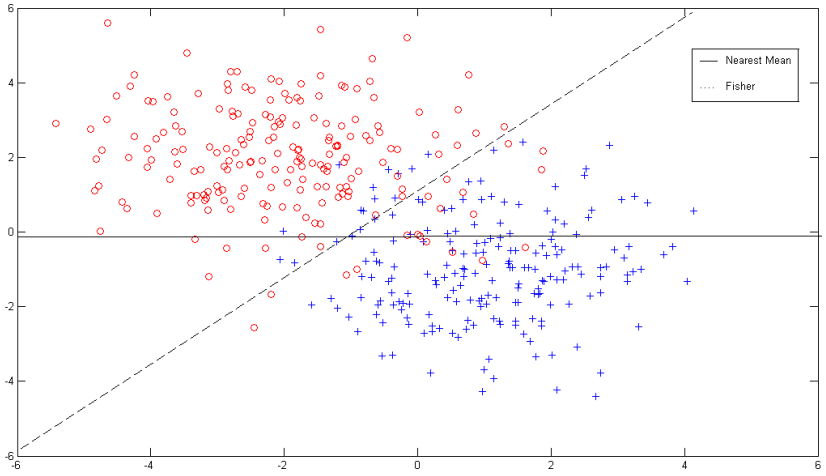

Figura 4.1: Comparación del argoritmo de clasificación basado en media mas cercana y fisher lineal.

de distancia a un vector media, ilustrando las ideas descritas en la anterior sección.

El clasificador de media más cercana es el clasificador óptimo en el caso de que las clases tengan la misma distribución extrictamente decreciente y simétrica alrededor de la media, y la misma varianza en todas las direcciones. Este es el caso de la distribución Gaussiana o normal. Además, si las medias de las poblaciones de las clases son conocidas, y no tienen que ser estimadas a través de las medias muestrales (4.17), el clasificador de media más cercana coinicide con el clasificador de Bayes (Fukunaga, 1990).

4.2.2. Clasificador linear de Fisher

El clasificador lineal de Fisher Fukunaga (1990); Raudys and Duin (1998) es una extensión del clasificador de media más cercana teniendo en cuenta la forma de la distribución de los vectores de características.

El clasificador de Fisher se define como:

$$f_{\mathrm{FL}}(\mathbf{x}') = (\mathbf{x}' - \mu_2)^T \mathbf{C}^{-1} (\mathbf{x}' - \mu_2) - (\mathbf{x}' - \mu_1)^T \mathbf{C}^{-1} (\mathbf{x}' - \mu_1) \tag{4.18}$$

donde $\mathbf{C}$ es la matriz de covarianza de la distribución de las muestras que se asume idéntica en ambas clases por razones computacionales. Un ejemplo de la ventaja de usar este clasificador sobre el clasificador de media más cercana se ilustra en la figura 4.1, donde se muestran las líneas separatorias estimadas para ambos clasificadores en el caso de dos clases bidimensionales distribuidas gaussianamente.

Si el número de muestras de entrenamiento es menor que la dimensión del espacio de características, la matriz de covarianza $\mathbf{C}$ es a una matriz singular, por lo que no puede ser invertida. En ese caso, se pueden seguir diferentes estrategias, como asumir restricciones en la matriz de covarianza (una matriz de covarianza proporcional a la identidad o diagonal) o reemplazar $\mathbf{C}^{-1}$ por su pseudo-inversa, por ejemplo usando una descomposición de valor singular para obtenerla (veáse (Fukunaga, 1990; Raudys and Duin, 1998)).

4.3. Selección del subconjunto de características

Una de las cuestiones de central importancia es la habilidad de un clasificador para generalizar, que no depende solo del clasificador elegido, sino también directamente de la construcción del vector de características. Los datos originales se pueden organizar de diferentes formas para construir estos vectores, siendo la primera cuestión a abordar la relación entre el tamaño del conjunto de muestras de entrenamiento n y la dimensión del espacio de características m.

La teoría de aprendizaje estadístico ofrece una primera respuesta esta cuestión, estableciendo la habilidad de un clasificador para agrupar de manera effectiva n puntos de un espacio de alta dimensionalidad en dos clases diferentes (Cover, 1965). Consideremos n puntos en un espacio de características m-dimensional. Se asume que los puntos estan *bien distribuidos*, de

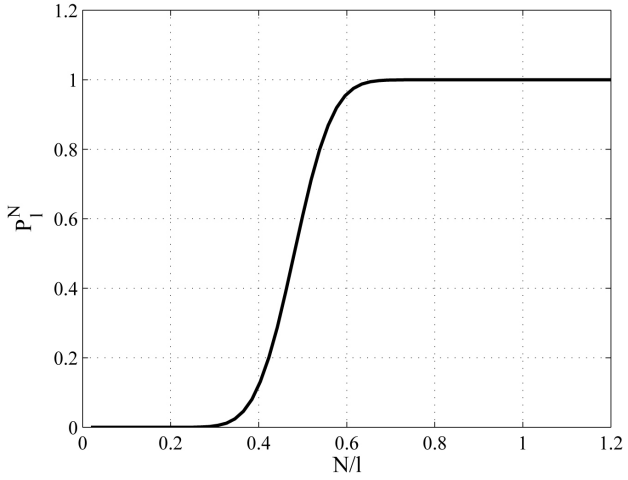

Figura 4.2: Probabilidad de agrupar n puntos en un espacio de características m-dimensional en 2 clases linealmente separables.

manera que no existe ningún subconjunto de n-1 puntos que se situen en un hiperplano de dimensión menor $m-1$. El número $O(l,N)$ de agrupaciones que pueden ser formadas por hiperplanos $m-1$-dimensionales para separar los n puntos en dos clases viene dado por (Cover, 1965):

$$O(n,m) = 2\sum_{i=0}^{m} \binom{n-1}{i} \tag{4.19}$$

donde

$$\binom{n-1}{i} = \frac{(n-1)!}{(n-1-i)!i!} \tag{4.20}$$

Por lo tanto, la probabilidad de agrupar n puntos en un espacio de características m-dimensional en 2 clases linealmente separables será:

$$P_n^m = \frac{O(n,m)}{2^n} = \begin{cases} \frac{1}{2^{n-1}} \sum_{i=0}^{m} \binom{n-1}{i} & n > m+1 \\ 1 & n \leq m+1 \end{cases} \tag{4.21}$$

La figura 4.2 muestra la probabilidad P_n^m como una función de m/n. Para espacios de características de baja dimension ($m/n < 0{,}3$), la probabilidad de separación de clases P_n^m es prácticamente cero, esto es, los clasificadores lineales tienen un bajo rendimiento discerniendo entre dos clases. Sin embargo, cuando se aumenta la dimensión del espacio de características, la probabilidad de que el conjunto de n puntos sea separable se aproxima a la unidad. Así, como intuitivamente parece claro, añadir información al vector de características mejora la separabilidad de clases para el caso de clasificador lineal.

Por otro lado, el hecho de que un cociente m/n pequeño produce bajos valores para probabilidad de separabilidad puede resolverse para un número fijo n de puntos a clasificar mediante un mapeo a un espacio de dimensión mayor. Esto puede conseguirse a través del uso de *kernels*, que involucran el uso de un producto interno no lineal 4.1, convirtiendo el espacio de características en otro de dimensión mayor donde P_n^m aumente, y un clasificador lineal pueda operar satisfactoriamente.

A la vista de la figura 4.2 y de la ecuación 4.21 es natural pensar que un aumento de la dimensión del espacio de características será siempre beneficioso para discriminar entre dos clases. Paradójicamente, ocurre exactamente lo contrario, dando lugar al problema conocido como la *maldición de la dimensionalidad*(en inglés, course of dimensionality) o el fenómeno del máximo (en inglés, peaking phenomenon). Este problema se describe como una reducción de la eficacia de un clasificador al añadir nuevas características a los vectores de entrenamiento cuando el número de estos es relativamente pequeño en comparación con número de características. El problema radica en que para definir un clasificador en un espacio de características de alta dimensionalidad es necesario estimar un numero de parámetros comparable a la dimensión del espacio. Por ejemplo, en el caso de un clasificador lineal, será necesario estimar $m + 1$ parámetros en un espacio de características m-dimensional (ver (4.15). Por lo tanto, aunque el clasificador separe los datos de entrenamiento satisfactoriamente, la fiabilidad en la estimación de los parámetros del clasificador será baja, ya que se estimarán muchos parámetros con un número muy reducido de vectores de entrenamiento. El clasificador construido con esta limitación tendrá, por consiguiente, una baja capacidad de generalización.

El problema de la maldición de la dimensionalidad justifica el uso de técnicas de reducción de la dimensionalidad del espacio de características, cuando

el numero de características usadas para diseñar el clasificador es mucho mayor que el número de vectores de entrenamiento disponibles. Aunque este es el principal motivo para hacerlo, existen otras motivaciones adicionales para reducir la dimensión del espacio de características hasta un mínimo razonable:

- la reducción del coste computacional de los algoritmos de entrenamiento y test
- eliminación de la correlación entre características
- selección de las características mas relevantes para la clasificación.

Así, aunque el problema de la maldición no exista, es natural suponer que el uso de un subconjunto de características con mejor capacidad de discriminación entre clases, optimizarán tanto el coste computacional del algoritmo de clasificación como el rendimiento del mismo.

Examinaremos dos métodos generales para la obtención del subconjunto de características: los métodos de filtro y los métodos de envoltura y estudiaremos algunos ejemplos concretos para ellos.

4.3.1. Métodos de filtro

El primer enfoque que examinaremos para la obtención del subconjunto de características para construir los vectores de entrenamiento y test, introduce un proceso independiente con este fin, que ocurre antes de la categorización del vector de características. Por esta razón, (John et˜al., 1994) los denominaron métodos de filtro, ya que mediante ellos se descartan los atributos irrelevantes para la clasificación antes de que ésta tenga lugar. Este paso de preprocesamiento de los datos usa aspectos generales del conjunto de entrenamiento para seleccionar o extraer unas características y excluir otras. De esta manera, los métodos de filtro no dependen del algoritmo de clasificación y podrán ser combinados con cualquiera de estos algoritmos, sin más que usar el subconjunto de características obtenido mediante filtrado para clasificación.

El método de filtro más sencillo que usaremos para reducir la dimensión del espacio de características en el caso de imagenes médicas será el de reducir el tamaño de las imágenes mediante *subsampling*. Si la compresión de las imagenes no es elevada, este es un método eficaz de disminuir la dimensión del espacio de caracteristicas, obteniendose un subconjunto de caracteristicas que recoge prácticamente la misma información que los datos originales. En este caso, no se seleccionan aquellas características relevantes para la clasificación, corriendose el riesgo de eliminar información importante para diseñar el clasificador. A continuación mostraremos ejemplos en los que se pretenede evitar este problema.

Máscara

Una manera de seleccionar los datos interesantes para la clasificación puede ser la construcción de una máscara binaria que extraiga de las imágenes, aquellos vóxels cuya intensidad sobrepase un límite prefijado. Una razón evidente para usar este método es que, fijando el umbral de selección adecuadamente, permitirá seleccionar aquellos voxeles de la imagen tomográfica que pertenzcan al cerebro, descartando todas aquellas regiones que quedan fuera del cerebro cuya intensidad es muy baja y no contienen información útil para la clasificación. La definición exacta parte del cálculo de la imagen media del conjunto de datos. Considerando que tenemos un conjunto de imagenes cerebrales $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, ..., \mathbf{\Gamma}_n$, se define la imagen media como:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \mathbf{\Gamma}_i \tag{4.22}$$

donde $\{\mathbf{\Gamma}_i \in \mathbb{R}^m\}$ es la imagen i muestral. Se define la máscara $\mathbf{E}$ binaria a partir de los valores ϵ_j:

$$\epsilon_j = \begin{cases} 1 & \text{si} \quad (\mu_j > t) \\ 0 & \text{si} \quad (\mu_j \leq t) \end{cases} \quad j = 1, ..., m \tag{4.23}$$

donde t es un umbral de intensidad fijado a priori. Con estos valores se puede contruir la matriz de máscara $\mathbf{E}$ como:

$$\mathbf{E} = \text{diag}\left(\epsilon_1, \epsilon_2, ..., \epsilon_m\right) \tag{4.24}$$

De esta manera, la aplicación de la máscara $\mathbf{E}$ sobre un vector imagen $\mathbf{\Gamma}_i$:

$$\hat{\mathbf{\Gamma}}_i^T = \mathbf{\Gamma}_i^T \mathbf{E} \tag{4.25}$$

producirá un nuevo vector columna $\hat{\mathbf{\Gamma}}_i$. Este nuevo vector contendrá únicamente la información de aquellos voxels cuya intensidad promedio en el conjunto de muestras supere un valor fijado por t. Si t es pequeño se obtendrá una mascara que seleccione el interior del cerebro, y aumentando paulatinamente el valor de t, se irán descartando aquellas regiones en las que la intensidad promedio no sea muy alta. Este segundo caso queda reflejado en la figura 4.3, y puede introducir mejoras para nuestros intereses, ya que las regiones cuya intensidad promedio sea baja, tanto en las imágenes de pacientes afectados por el Alzheimer como para los pacientes control, son regiones que no tienen interés para el diagnóstico del Alzheimer. Esto ocurre tanto en imágenes PET como SPECT ya que, en ambos casos, las imágenes afectadas por la enfermedad presentan unos niveles de intensidad menores que en las normales (Goethals et˜al., 2002), por lo que su valor promedio será intermedio y no bajo.

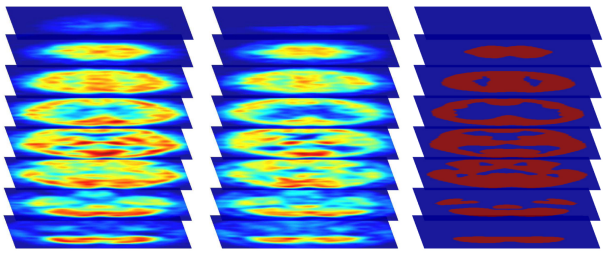

Figura 4.3: Secciones transversales de: *Columna izquierda*: Un paciente normal. *Columna central*: Un paciente DTA. *Columna derecha*: Máscara

Transformación lineal

La siguiente manera de reducir la dimensionalidad de los vectores de imágenes $\mathbf{\Gamma}_i^T = (\Gamma_1, \Gamma_2, ..., \Gamma_m)_i$ será aplicandoles una tranformación lineal del tipo:

$$\mathbf{z}_i = \mathbf{\Gamma}_i^T \mathbf{P} \tag{4.26}$$

donde $\mathbf{P}$ es una matriz $q \times m$ formada por p vectores columna $\mathbf{p}_i = (p_1, p_2, ..., p_m)_i$: $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_q]$ y $\mathbf{z}_i = (z_1, z_2, ..., z_q)_i$ es el vector de características reducido. Expandiendo los vectores de la ecuación anterior en coordenadas:

$$\begin{pmatrix} (z_1, z_2, ..., z_q)_1 \\ (z_1, z_2, ..., z_q)_2 \\ \vdots \\ (z_1, z_2, ..., z_q)_n \end{pmatrix} = \begin{pmatrix} (\Gamma_1, \Gamma_2, ..., \Gamma_m)_1 \\ (\Gamma_1, \Gamma_2, ..., \Gamma_m)_2 \\ \vdots \\ (\Gamma_1, \Gamma_2, ..., \Gamma_m)_n \end{pmatrix} \left(\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix}_1, \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix}_2, ^{\cdots}, \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix}_q \right) \tag{4.27}$$

Si $q = m$, entonces la ecuación (4.26) solo supondrá una transformación lineal de los datos, asi que en general, estaremos interesados en transformaciones con $q << m$.

Esta clase de métodos de filtro se basa en al extracción de características y no en la selección, consiguiendo un subconjunto de características relevantes transformando los datos originales en un nuevo conjunto. El hecho de trabajar en espacios Euclídeos, permite dar una interpretación geométrica sencilla a este tipo de tranformaciones (4.26). Tomando una base cartesiana de $\mathbb{R}^m$, es decir:

$$\begin{aligned} \mathbf{u}_1 &= (1, 0, ..., 0) \\ \mathbf{u}_2 &= (0, 1, ..., 0) \\ &\vdots \\ \mathbf{u}_m &= (0, 0, ..., 1) \end{aligned} \tag{4.28}$$

donde $\mathbf{u}_i \in \mathbb{R}^m$, $i = 1, .., m$, podemos expresar detalladamente el vector imagen $\mathbf{\Gamma}$ como:

$$\mathbf{\Gamma} = \sum_{i=1}^{m} \Gamma_i \mathbf{u}_i \tag{4.29}$$

que es su expansión en coordenadas. A menudo omitiremos que estamos trabajando en esta base cartesiana, y expresaremos el vector $\mathbf{\Gamma}$ simplemente a través de sus coordenadas en esta base $\mathbf{\Gamma} = (\Gamma_1, \Gamma_2, ..., \Gamma_m)$. Si queremos expresar el vector imagen $\mathbf{\Gamma}$ en otra base $\{\mathbf{p}_i\}$ de $\mathbb{R}^m$, entonces su representación en esta nueva base será:

$$\mathbf{\Gamma} = \sum_{i=1}^{m} x_i \mathbf{p}_i \tag{4.30}$$

donde las coordenadas x_i del vector $\mathbf{\Gamma}$ en la nueva base se relacionan con las anteriores en la base cartesiana mediante una tranformación del tipo (4.26), con $q = m$. Por lo tanto, si $q < m$, entonces (4.26) no es más que la expresión de las coordenadas del vector $\mathbf{\Gamma}$ en el subespacio engendrado por los vectores $\mathbf{p}_i \in \mathbb{R}^m$, $i = 1, ..., q$. En otras palabras, la tranformación (4.26) se puede visualizar como la *proyección del vector* $\mathbf{\Gamma}$ *en un subespacio*, que queda definido por los vectores $\mathbf{p}_i$. El objetivo será encontrar un subespacio del espacio de características $\mathcal{H}$ de dimensión menor, en el que los vectores originales queden fielmente representados, o al menos sus atributos más relevantes para la clasificación. Se conseguirán diferentes realizaciones de esta disminución de la dimensionalidad en función de la elección de los vectores $\mathbf{p}_i$ que definen la matriz $\mathbf{P}$.

Un ejemplo de este tipo de transformaciones lo constituye la máscara de la sección anterior, que se puede considerar como un caso particular de transformación lineal. La ecuación (4.25) es de la forma (4.26), identificando $\mathbf{E}$ con $\mathbf{P}$. Puesto que la matriz $\mathbf{E}$ es diagonal y binaria, la interpretación geométrica de esta transformación es sencilla: mediante la máscara se selecciona un subconjunto de vectores de la base (4.28) para representar la imagen, descartándose aquellos que determina la ecuación (4.24) y sin cambiar de base.

La técnica estadística de análisis de componentes principales (Jolliffe, 2002), constituye el siguiente ejemplo más conocido de criterio para definir la matriz $\mathbf{P}$, aunque no se presente tradicionalmente de esta manera. En este análisis se generan combinaciones lineales de los elementos originales, cuya matriz de tranformación esta formada por vectores que son ortogonales en el espacio original. Sin embargo es necesario cierto criterio para seleccionar un subconjunto de los datos transformados, que permite reducir la dimensionalidad del espacio de características original, por lo que la transfomación no toma la forma de (4.26). Veremos en 9.2.2 una explicación detallada de ésto, mientras que en 9.4 estudiaremos una variante mas relacionada con (4.26).

Empíricamente, las componentes principales han logrado reducir la dimensionalidad de una amplia variedad de problemas de aprendizaje. (Blum and Langley, 1997) describen las garantías teóricas de los métodos de esta forma, cuando la función objetivo es una intersección de halfspaces y las muestras son elegidos de una distribución sufisuficientemente benigna. El método de análisis de componentes independientes (Comon, 1994), relacionado con el anterior, incorpora ideas similares, pero insistiendo en la independencia de las nuevas características en lugar de en su ortogonalidad.

En las secciones 9.2 y 10.2, explicaremos detalladamente cómo funcionan estos métodos de filtro y haremos uso de ellos combinandolos con otras ideas.

4.3.2. Métodos de Envoltura

El segundo enfoque genérico para la selección de características también se produce fuera del proceso de clasificación, aunque utilizando la clasificación como subrutina, en lugar de como un postprocesador. Por esta razón, (John et˜al., 1994) se refieren a estos como métodos de envoltura (en inglés wrapping methods (Kohavi and John, 1997)). El algoritmo típico de este tipo busca subconjuntos en el espacio de características que produzcan resultados óptimos en la clasificación, ejecutando internamente la función clasificadora en cada alternativa. Tras este proceso, la selección de características se lleva a cabo usando como criterio la estimación de la precisión del clasificador, seleccionándose aquellas características que produjeron mejores resultados, y descartando el resto. En realidad, la los métodos de envolutra tienen una larga historia dentro de la literatura sobre estadística y reconocimiento de

patrones (por ejemplo, (Devijver and Kittler, 1982)), donde el problema de la selección de características ha sido un tema de investigación activo durante mucho tiempo, pero su uso dentro del aprendizaje automático es relativamente reciente.

El argumento general a favor de los métodos de envotura es que al usar la clasificación internamente, se obtendrá una estimación mejor de la precisión en ese subconjunto que si se usa una medida separada con otro sesgo. Por ejemplo, tanto (Doak, 1992) y (John et˜al., 1994) defienden la utilización de un método de envoltura para mejorar el comportamiento de inducción de los árboles de decisión, o (John et˜al., 1994) que presenta estudios comparativos sobre los efectos de usar métodos de filtro frente a métodos de envoltura.

Los métodos de envoltura puede proporcionar soluciones más precisas que los métodos de filtro al problema de la selección de características (Kohavi and John, 1997). Sin embargo, el principal inconveniente de los métodos de envoltorio frente a los métodos de filtro es el del coste computacional del primero, resultado de llamar al algoritmo de clasificación en cada conjunto de características consideradas, que debe ser evaluado utilizando un subconjunto excluído del proceso.

4.4. Parametros de valoración del rendimiento de un clasificador

El objetivo último de un clasificador consiste en asignar correctamente una etiqueta a un patrón de test. En el caso de clasificación binaria, en la que los datos estan divididos entre etiquetas positivas o negativas, existirán dos posibles errores que el clasificador puede cometer: clasificar como positivo un patrón que en realidad era negativo o vicesversa. Estas posibilidades vienen reflejadas en la tabla 4.1, donde se compara el resultado del test con la etiqueta original, dando lugar a PV (Positivo Verdadero) cuando los dos coinciden en valor positivo, PF (Positivo Falso) cuando el test da positivo mientras la etiqueta original era negativa, NF (Negativo Falso) cuando el test da negativo mientras la etiqueta original era positiva y NV (Negativo Verdadero) cuando los dos coinciden en valor negativo.

Tabla 4.1: Posibles resultados del test en función de la etiqueta

		Etiqueta		
		Positiva	Negativa	
Test	Positivo	PV	PF	→ *Valor predictivo positivo*
	Negativo	NF	NV	→ *Valor predictivo negativo*
		↓	↓	
		Sensibilidad	*Especificidad*	

La capacidad de un clasificador para detectar los positivos verdaderos se medirá a través de la sensibilidad, que se define como:

$$\text{Sensibilidad} = \frac{\text{número de } PV}{\text{número de } PV + NF} \tag{4.31}$$

de manera que una sensibilidad del 100 % corresponderá a un clasificador que es capaz de detectar todos los pacientes etiquetados como positivos como tales. Por lo tanto, si un clasificador con alta sensibilidad da un resultado negativo, éste será muy fiable, lo que puede ser usado para descartar la enfermedad. La sensibilidad esta relacionada con el error de tipo I en inferencia estadística, que consiste en rechazar la hipótesis nula cuando en realidad es cierta.

Por otro lado, la capacidad para detectar negativos verdaderos vendrá dada por la especificidad, definida como:

$$\text{Especificidad} = \frac{\text{número de } NV}{\text{número de } NV + PF} \tag{4.32}$$

permitiendo que un clasificador con alta especificidad sea muy fiable a la hora de confirmar la enfermedad, ya que raramente producirá un resultado positivo que en realidad sea falso. La especificidad esta relacionada con el error de tipo II donde se acepta la hipótesis nula cuando en realidad es falsa.

Sin embargo, tanto altos valores de la sensibilidad como de la especificidad no tienen porque corresponder a un clasificador preciso. Se define la precisión como:

$$\text{Precisión} = \frac{\text{número de } PV + NV}{\text{número de } PV + PF + NF + NV} \tag{4.33}$$

Puede ocurrir que un clasificador tenga valores cercanos al 100 % de sensibilidad y cercanos al 0 % de especificidad. Este clasificador no tendrá capacidad de discernir entre las clases, ya que será un clasificador que tome cualquier patrón como positivo. Esto es equivalente a una clasificación al azar, ya que su precisión rondará el 50 % para una muestra sin preponderancia de ninguna de las dos clases. El clasificador deseable será aquel que tenga valores altos de sensibilidad, especificidad y precisión simultáneamente, y no solo de alguno de ellos por separado.

Otros parámetros que pueden resultar interesantes son los valores predictivos. Éstos hacen referencia a la validez de un resultado de clasificación positivo/negativo (valor predictivo positivo/negativo). Se podrá confiar más en un resultado positivo de un clasificador con un vpp alto que uno con un vpp menor. Sin embargo, los valores predictivos dependen de la preponderancia de las clases, denominada pervalencia, término de epidemiología que determina la proporción de individuos de una población que, en este caso, padece la enfermedad. Si el conjunto de test no tiene igual número de positivos que de negativos, habrán de usarse las fracciones de probabilidad positiva o negativa (fpp/fpn):

$$\text{fpp} = \frac{\text{sensibilidad}}{1 - \text{especificidad}} \tag{4.34}$$

y

$$\text{fpn} = \frac{1 - \text{sensibilidad}}{\text{especificidad}} \tag{4.35}$$

que no dependen de la prevalencia.

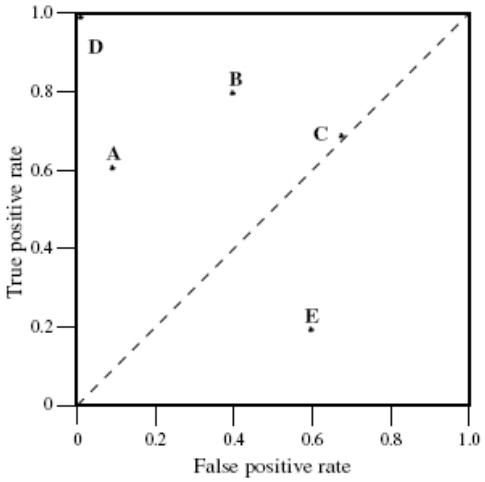

Figura 4.4: Representación en el espacio ROC

4.4.1. Curva ROC

A menudo será interesante valorar cómo se modifica el rendimiento de un clasificador al moficar algun parámetro, ya sea del clasificador o en lac definición de algun paso anterior. Para este análisis será útil la representación en el espacio ROC (receiver operating characteristic (ROC) (Fawcett, 2006)), que no es más que una representación bidimensional de la tasa de positivos verdaderos (sensibilidad) frente a la tasa de falsos positivos (1-especificidad).

En este espacio (vease figura 4.4), la mejor predicción correponderá con un punto cercano a la esquina superior izquierda (D), representando sensibilidad del 100 % (ningún negativo falso) y especificidad 100 % (ningún positivo falso), que producirá tambien una precisión del 100 %. El punto D se llama tambien clasificación perfecta. Una clasificación completamente aleatoria daría un punto C en la línea diagonal (llamada la línea de no-discriminación) de la parte inferior izquierda a la parte superior derecha. Por debajo de esta línea estarían resultados peores que la estimación al azar E. Los puntos A y B representan clasificaciones adecuadas, la primera con mayor sensibilidad que espeficicidad y la segunda con una especificidad mayor que la sensibilidad.

4.5. Métodos de Validación Cruzada

La Validación cruzada, a veces denominada estimación por rotación (Kohavi, 1995; A. and Kittler, 1982), es una técnica muy popular para evaluar cómo el resultado de un análisis estadístico tiene mayor o menor capacidad de generalización sobre un conjunto independiente de datos. Se usa preferentemente en aplicaciones donde el objetivo último es la predicción , y se emplea para estimar cómo de preciso es un modelo predictivo en la práctica. Una ronda de validación cruzada consiste en dividir el conjunto muestral en subconjuntos complementarios, y realizar el análisis en un subconjunto (llamado el conjunto de entrenamiento), y validar éste análisis en otro subconjunto (llamado subconjunto de validación o de test). Para reducir la variabilidad en la evaluación global de la generalización, se levan a cabo múltiples rondas de validación cruzada usando distintas particiones, y los resultados de la validación son promediados sobre todas las rondas realizadas.

En los métodos de validación cruzada, el conjunto de test no supone un test 'real', ya que la etiqueta de los elementos del conjunto de test es conocida. De esta manera, se puede comparar el resultado del test con la etiqueta original, y determinar si se trata de un PV, PF, NV, o NF. Una vez el proceso de itera sobre cada partición, se puede calcular cualquiera de las cantidades definidas en la sección anterior.

La teoría de la Validación Cruzada fue originalmente desarrollada por Seymour Geisser y es de radical importancia para vigilar la posible presencia del error estadístico de tipo III en cualquier proceso de decisión. Este tipo de error consiste en rechazar la hipótesis nula de manera correcta por razones erróneas que ocurre, por ejemplo, cuando el tamaño muestral es limitado. De entre los métodos de validación destacan:

4.5.1. Validación por sub-muestreo aleatorio repetido

Este método aleatoriamente divide aleatoriamente el conjunto de datos en dos conjuntos de entrenamiento y validación. Para cada división, el clasificador es entrenado con el conjunto de entrenamiento y validado sobre los datos restantes. Los resultados de cada división son promediados. La ventaja de este método es que la proporción de la división entrenamiento/validación

no depende del número de iteraciones pero la principal desventaja del mismo es que algunas muestras puede nunca ser seleccionadas en el subconjunto de validación, mientras que otras pueden ser seleccionadas más de una vez. Es decir, los conjuntos de validación pueden solaparse. Este método también exhibe la variación Monte Carlo, esto es los resultados variarán si el análisis es repetido con diferentes conjuntos aleatorios. Una variante de esta aproximación genera muestras aleatorias de tal forma que el valor de respuesta medio es igual en los subconjuntos de entrenamiento y test. Esto es particularmente útil cuando el conjunto muestral contiene una representación no balanceada en las respuestas de las muestras.

4.5.2. Validación Cruzada *K*-pliegues

En la validación K-pliegues, el conjunto muestral original es divido en K subconjuntos (Breiman et˜al., 1984). De los K subconjuntos, uno de ellos se destina a validación para testar el modelo y los $K-1$ restantes se usan como conjunto de entrenamiento. Después la validación cruzada se repite K veces (los pliegues), con cada uno de los K subconjuntos usados una vez como datos de validación. Los K resultados de cada pliegue son promediados (ó combinados) para producir una única estimación.

La ventaja de este método sobre el anterior es que todas las observaciones se usan tanto para entrenamiento como para test y cada observación se usa para validación solo una vez. Este método se suele usar cuando el número de elementos de la muestra es muy grande, o cuando los algoritmos de clasificación son computacionalmente costos, de manera que se tiene control sobre el número de veces que se itera la validación a través del número K. En este caso también puede considerarse que en cada pliegue se contenga la misma proporción de etiquetas o respuestas.

4.5.3. Validación dejar uno fuera

Como el nombre indica, la validación dejar uno fuera consiste es utilizar una sola muestra como observación de test y las restantes observaciones como entrenamiento (Raudys and Jain, 1991). Este proceso se repite N veces de manera que todas las muestras son usadas una vez como observaciones para

la validación, por lo que se considera un caso particular del anterior método de validación. La ventaja del método de dejar uno fuera es que el conjunto de entrenamiento es lo más grande que la muestra permite, aumentando la estadística en la estimación de los parámetros del calsificador. Por contra, este método puede ser computacionalmente costoso dado el gran número de veces que se repite el proceso de validación, si el número de elementos de la muestra es grande.

En el caso de diagnóstico médico, éste será a menudo el método de validación elegido, ya que el número de pacientes de una base de datos habitual es suficientemente reducido para que el uso de la validación dejar uno fuera no suponga un gran coste computacional.

CAPÍTULO 5

Máquinas de Soporte Vectorial

Las máquinas de Soporte Vectorial se usan ampliamente para el reconocimiento de patrones en un gran numero de aplicaciones por su habilidad para aprender de datos experimentales(Burges, 1998; Joachims, 1998). SVM ha atraído recientemente la atención de la comunidad dedicada al reconocimiento de patrones debido a la cantidad de méritos derivados de la Teoría del Aprendizaje Estadístico (Vapnik, 1995, 1998) desarrollada por Vladimir Vapnik en AT&T. Estas técnicas se han usado en una gran cantidad de aplicaciones incluyendo la detección de actividad vocal (VAD) (Enqing et˜al., 2002a,b; Qi et˜al., 2004; Ramírez et˜al., 2006b,a; Yélamos et˜al., 2006), recuperación de imágenes basadas en contenido (Tao et˜al., 2006), clasificación de texturas (Kim et˜al., 2002) y diagnosis de imágenes médicas (Fung and Stoeckel, 2007; Kalatzis et˜al., 2003; Illán et˜al., 2009; López et˜al., 2009). La razón de este interés reside en que SVM son mucho mas efectivos que otros clasificadores convencionales.

5.1. SVM Lineal

Una de las mayores ventajas de los clasificadores lineales es su simplicidad y atractivo computacional. En este apartado supondremos que todos los vectores de características de las clases disponibles pueden clasificarse correctamente usando un clasificador lineal. Más adelante nos centraremos en problemas más genéricos donde los clasificadores lineales no pueden clasificar correctamente todos los vectores, y trataremos de buscar modos de diseñar un clasificador óptimo lineal adoptando un criterio de optimización apropiado.

5.1.1. Clases linealmente separables

Los clasificadores lineales definen hipersuperficies o hiperplanos de decisión en espacios multidimensionales, esto es:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0, \tag{5.1}$$

donde $\boldsymbol{w}$ se conoce como vector de peso y w_0 como el umbral. El vector de pesos $\boldsymbol{w}$ es ortogonal al hiperplano de decisión y la tarea de optimización consiste en encontrar el conjunto de parámetros w_i, i= 1, ..., N, que definen el hiperplano de decisión.

Sea $\boldsymbol{x}_i$, i=1, 2, ..., l, los vectores de características del conjunto de entrenamiento, X. Estos pertenecen a una de las dos clases, ω_1 o ω_2. Si las clases son linealmente separables el objetivo sería diseñar un hiperplano que separe completamente todos los vectores de entrenamiento. Este hiperplano no es único y el proceso de selección se centra en maximizar la *generelización* del clasificador. De entre los posibles criterios, se selecciona el de la búsqueda del hiperplano que deje el máximo margen entre clases (conocido como el hiperplano de margen maximal), ya que hace máxima la separación entre clases.

El objetivo será pues, buscar la dirección que da el máximo margen posible. Puesto que no queremos dar preferencia a ninguna de las dos clases, entonces es razonable elegir para todas las direcciones el hiperplano que dista lo mismo respectivamente de los puntos más cercanos en ω_1 y ω_2. Puesto que la distancia desde un punto $\boldsymbol{x}$ el hiperplano viene dada por:

$$z = \frac{|g(\boldsymbol{x})|}{||\boldsymbol{w}||} \tag{5.2}$$

escalamos $\boldsymbol{w}$ y w_0 de modo que el valor de $g(\boldsymbol{x})$ en los puntos más cercanos sea +1 para el punto más cercano en ω_1 y -1 para el punto más cercano en ω_2 . Esto reduce el problema de optimización a maximizar el margen:

$$\frac{1}{\|\boldsymbol{w}\|} + \frac{1}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|} \tag{5.3}$$

sujeto a las condiciones:

$$\begin{array}{ll} \boldsymbol{w}^T\boldsymbol{x} + w_0 \geq 1, & \forall \boldsymbol{x} \in \omega_1, \\ \boldsymbol{w}^T\boldsymbol{x} + w_0 \leq 1, & \forall \boldsymbol{x} \in \omega_2, \end{array} \tag{5.4}$$

Para cada $\boldsymbol{x}_i$ denotamos el correspondiente indicador de clase y_i (+1 para ω_1 y -1 para ω_2). Nuestra tarea puede resumirse así: Calcular los parámetros $\boldsymbol{w}$ y w_0 del hiperplano de manera que se minimice la siguiente expresión

$$\mathbf{J}(\boldsymbol{w}) \equiv \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{5.5}$$

sujeto a

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 \quad \text{para} \quad i = 1, 2, ..., N \tag{5.6}$$

Obviamente, minimizando la norma el margen se hace mínimo. Esto es una tarea de optimización (cuadrática) no lineal sujeto a un conjunto de restricciones de inecuaciones lineales. Las condiciones de Karush-Kuhn-Tucker (KKT) establecen que ha de cumplirse lo siguiente:

$$\frac{\partial}{\partial \boldsymbol{w}}\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0} \tag{5.7}$$

$$\frac{\partial}{\partial w_0}\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = 0 \tag{5.8}$$

$$\lambda_i \geq 0 \quad \text{para} \quad i = 1, 2, ..., N \tag{5.9}$$

$$\lambda_i \left[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 \right] = 0 \quad \text{para} \quad i = 1, 2, ..., N \tag{5.10}$$

donde λ es el vector de multiplicadores de Lagrange, λ_i, y $\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda})$ es la función Lagrangiana definida como

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{N} \lambda_i \left[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 \right] \tag{5.11}$$

Combinando 5.11 con 5.7 y 5.8 resulta

$$\boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i \tag{5.12}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{5.13}$$

Los multiplicadores de Lagrange pueden ser cero o positivos. Por tanto, el vector de parámetros $\boldsymbol{w}$ de la solución óptima es una combinación lineal de $N_s \leq N$ vectores características asociados a $\lambda_i \neq 0$, es decir,

$$\boldsymbol{w} = \sum_{i=1}^{N_s} \lambda_i y_i \boldsymbol{x}_i \tag{5.14}$$

Éstos se conocen como los *vectores de soporte* y el hiperplano clasificador óptimo *máquina de vectores de soporte* (SVM). Al igual que para el conjunto de restricciones en 5.10 para $\lambda_i \neq 0$, los vectores de soporte caen en uno de los dos hiperplanos

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 = \pm 1 \tag{5.15}$$

es decir, son los vectores de entrenamiento que están más cerca del clasificador lineal, y constituyen los elementos críticos del conjunto de entrenamiento.

Aunque $\boldsymbol{w}$ se da explícitamente, w_0 se puede obtener implícitamente por una de las condiciones 5.10. En la práctica, w_0 se calcula como un valor medio obtenido usando todas las condiciones de este tipo. Por otro lado, las propiedades de la función de coste 5.5 garantizan que la matriz Hessiana correspondiente es definida positiva. Además, las restricciones consisten en funciones lineales. Estas dos condiciones garantizan que cualquier mínimo local es también global y único. *El hiperplano clasificador de una máquina de vectores de soporte es único.*

Habiendo establecido todas estas propiedades interesantes del hiperplano óptimo de una máquina de vectores de soporte, el siguiente paso es el cálculo de los parámetros involucrados. Desde un punto de vista computacional esto no siempre es una tarea fácil, y existen numerosos algoritmos para ello. Se trata de un problema de la familia de programación convexa. Estos problemas se resuelven considerando la denominada *dualidad Lagrangiana*, y el problema puede formularse equivalentemente como sigue

Maximizar

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) \tag{5.16}$$

sujeto a

$$\boldsymbol{w} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i \tag{5.17}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{5.18}$$

$$\boldsymbol{\lambda} \geq \mathbf{0} \tag{5.19}$$

Las dos restricciones de igualdad son el resultado de igualar a cero el gradiente de la Lagrangiana, con respecto a $\boldsymbol{w}$ y w_0. Los vectores de características de entrenamiento aparecen en el problema mediante las restricciones de

igualdad y no mediante las inecuaciones, lo cual hace más fácil de manejar. Sustituyendo 5.17 y 5.18 en 5.16 y haciendo algunas operaciones llegamos a la tarea de optimización equivalente

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \tag{5.20}$$

sujeto a

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{5.21}$$

$$\boldsymbol{\lambda} \geq \mathbf{0} \tag{5.22}$$

Una vez que los multiplicadores de Lagrange han sido calculados, maximizando 5.20, el hiperplano óptimo se calcula vía 5.17 y w_0 como antes.

Observaciones:

- Además de ser una manera más cómoda, existe otra razón por la cual se opta por la formulación de 5.20 y 5.21. Los vectores de entrenamiento aparecen en parejas, en la forma de productos escalares. Esto es lo más interesante. *La función coste no depende explícitamente de la dimensionalidad del espacio de entrada.* Esta propiedad permite generalizaciones eficientes para el caso de clases linealmente no separables.
- Aunque el hiperplano óptimo resultante es único, no existe garantía de la unicidad de los multiplicadores de Lagrange asociados, λ_i. En resumen, la expansión de $\boldsymbol{w}$ en términos de vectores soporte en 5.17 puede no ser única, aunque el resultado final es único.

5.1.2. Clases linealmente no separables

En el caso de que las clases no sean separables, lo dicho anteriormente deja de ser válido. Cualquier intento de dibujar un hiperplano no conseguirá una

banda de separación de clases sin puntos dentro de ella, como era el caso de clases linealmente separables. Recordando que el margen se define como la distancia entre el par de hiperplanos paralelo descritos por

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 = \pm 1 \tag{5.23}$$

Los vectores de características de entrenamiento ahora pertenecen a una de las siguientes tres categorías

- Vectores que caen fuera de la banda y que son correctamente clasificados. Estos vectores cumplen con las restricciones en 5.5.
- Vectores que caen dentro de la banda y que son correctamente clasificados. Estos son los puntos rodeados por cuadrados en la Figura y satisfacen la inecuación

$$0 \leq y_i(\boldsymbol{w}^T\boldsymbol{x} + w_0) < 1 \tag{5.24}$$

- Vectores que son clasificados erróneamente. Éstos están rodeados por círculos y cumplen la inecuación

$$y_i(\boldsymbol{w}^T\boldsymbol{x} + w_0) < 0 \tag{5.25}$$

Estos tres casos se pueden tratar como un solo tipo de restricciones introduciendo un nuevo conjunto de variables

$$y_i(\boldsymbol{w}^T\boldsymbol{x} + w_0) \geq 1 - \xi_i \tag{5.26}$$

La primera categoría de datos corresponde con $\xi_i = 0$, la segunda con $0 < \xi_i \leq 1$ y la tercera con $\xi_i > 1$. Las variables ξ_i se conocen como *variables débiles*. La tarea de optimización se vuelve más complicada, aunque se basa en los mismos principios que antes. El objetivo ahora es hacer el margen tan grande como sea posible pero al mismo tiempo mantener la cantidad de puntos con $\xi_i \geq 0$ tan pequeña como sea posible. En términos matemáticos, esto es equivalente a minimizar la función de coste

$$J(\boldsymbol{w}, w_0, \xi) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\mathbf{I}(\xi_i), \tag{5.27}$$

donde $\boldsymbol{\xi}$ es el vector de parámetros ξ_i y

$$\mathbf{I}(\xi_i) = \begin{cases} 1, & \xi_i > 0 \\ 0, & \xi_i = 0 \end{cases} \tag{5.28}$$

El parámetro C es una constante positiva que controla la influencia relativa de los dos términos competitivos. Sin embargo, la optimización de arriba es difícil puesto que incluye una función discontinua $\mathbf{I}(\cdot)$. Como es común en casos así, se elige optimizar una función de coste estrechamente relacionada, y el objetivo se convierte en

Minimizar

$$J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i, \tag{5.29}$$

sujeto a

$$\begin{array}{ll} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i, & i = 1, 2, ..., N \\ \xi_i > 0, & i = 1, 2, ..., N \end{array} \tag{5.30}$$

El problema es de nuevo un problema de programación convexa, y la Lagrangiana correspondiente viene dada por

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 + \xi_i] \tag{5.31}$$

Realizando pasos similares a los del caso de clases separables, llegamos al siguiente problema de optimización equivalente

$$\max_{\lambda}(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j) \tag{5.32}$$

sujeto a

$$0 \leq \lambda_i \leq C, \text{para} \quad i = 1, 2, ..., N \tag{5.33}$$

$$\sum_{i=1}^{N}\lambda_i y_i = 0 \tag{5.34}$$

Observaciones:

- La única diferencia con el caso previamente considerado de clases linealmente separables está en la primera de las dos restricciones, donde es necesario limitar por arriba a los multiplicadores de Lagrange por C. El caso linealmente separable corresponde con $C \to \infty$. Las variables débiles ξ_i, y sus multiplicadores de Lagrange asociados, μ_i, no intervienen en el problema explícitamente. Su presencia está reflejada indirectamente mediante C.

- En todo este estudio se ha considerado sólo el caso de clasificación con dos clases. En el caso de M-clases, se puede extender fácilmente mirando el problema como M problemas de dos clases. Para cada una de las clases, tratamos de diseñar una función discriminante óptima, $g_i(\boldsymbol{x})$, i=1,2,...,M, de modo que $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \forall j \neq i$ si $\boldsymbol{x} \in \omega_i$. Adoptando la metodología de SVM podemos diseñar las funciones discriminantes de modo que $g_i(\boldsymbol{x}) = 0$ sea el hiperplano óptimo para separar la clase ω_i de todas las demás, dado por supuesto que esto es posible. Así, la función lineal resultante dará $g_i(\boldsymbol{x}) > 0$ para $\boldsymbol{x} \in \omega_i$ y $g_i(\boldsymbol{x}) < 0$ en caso contrario. La clasificación se consigue de acuerdo a la siguiente regla

$$Asignar\ \boldsymbol{x}\ a\ \omega_i\ si\ i = \arg_k \max \{g_k(\boldsymbol{x})\}$$

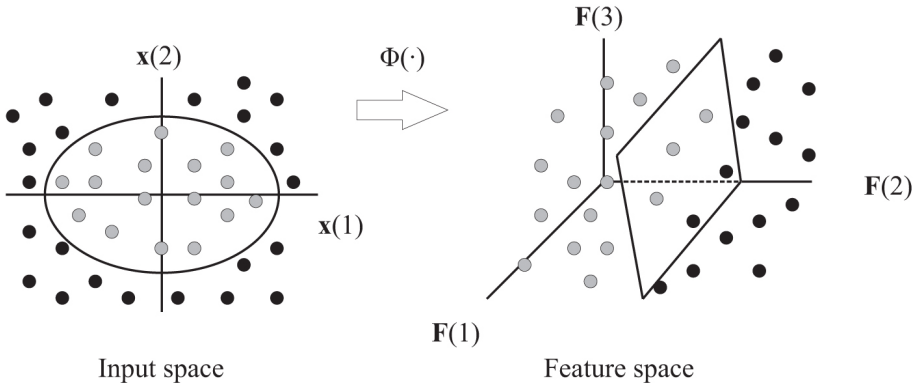

Figura 5.1: Mapeo del espacio de características a otro espacio donde la frontera de separación se hace lineal

Esta técnica, sin embargo, puede conducir a regiones indeterminadas, donde más de un $g_i(\boldsymbol{x})$ es positivo. Otra aproximación es extender la formulación matemática de SVM de dos clases al problema de M clases.

5.2. SVM No Lineal

En el apartado anterior se discutieron las máquinas de vectores de soporte (SVM) como una metodología óptima de diseño de un clasificador lineal. Asumimos ahora que existe un mapeo

$$\boldsymbol{x} \in \mathbb{R}^l \rightarrow \boldsymbol{y} \in \mathbb{R}^k$$

desde el espacio de entrada a un espacio k-dimensional, donde las clases se pueden separar satisfactoriamente por un hiperplano lineal. Recordamos que los vectores de características participan por pares mediante la operación del producto interno. También, una vez que el hiperplano óptimo $(\boldsymbol{w}, w_0)$ se ha calculado, la clasificación se realiza según si el signo de

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=0}^{N_s} \lambda_i y_i \boldsymbol{x}^T\boldsymbol{x} + w_0 \tag{5.35}$$

es positivo o negativo, donde N_s es el número de vectores de soporte. Así, una vez más, sólo el producto interno entra en escena. Si el diseño se va a llevar a cabo en el espacio k-dimensional, la única diferencia es que los vectores involucrados estarán en los mapeos k-dimensionales del vector de características original. Una simple ojeada a esto nos llevaría a la conclusión de que ahora la complejidad es mucho mayor, puesto que, habitualmente, k es mucho más alto que la dimensión l del espacio de entrada, para poder hacer las clases linealmente separables.

Una vez que el kernel adecuado se adopta, que implícitamente define un mapeo a un espacio de dimensión mayor, la tarea de clasificación se convierte en

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) \tag{5.36}$$

sujeto a

$$0 \leq \lambda_i \leq C, \quad i = 1, 2, ..., N \tag{5.37}$$

$$\sum_i \lambda_i y_i = 0 \tag{5.38}$$

y el clasificador lineal resultante es

$$Asignar\ \boldsymbol{x}\ a\ \omega_1(\omega_2)\ si\ g(\boldsymbol{x}) = \sum_{i=1}^{N_s} \lambda_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + w_0 > (<) 0$$

La Figura muestra la arquitectura correspondiente. El número de nodos viene determinado por el número de vectores de soporte N_s. Los nodos realizan el producto interno entre el mapeo de $\boldsymbol{x}$ y los correspondientes mapeos de los vectores de soporte en el espacio de dimensión mayor, mediante la operación de kernel.

Observaciones:

- Una característica notable de las máquinas de vectores de soporte es que la complejidad computacional es independiente de la dimensionalidad

del espacio kernel, donde las características de entrada son mapeadas. Así, uno diseña en un espacio de dimensión alta sin tener que adoptar modelos explícitos usando una gran cantidad de parámetros.

- Una limitación importante de las máquinas de vectores de soporte es la alta carga computacional que se requiere, tanto durante el entrenamiento como durante el test. Para problemas con una cantidad relativamente pequeña de datos de entrenamiento, se puede usar cualquier algoritmo de optimización de propósito general. Sin embargo, para una gran cantidad de puntos de entrenamiento (del orden de unos miles), se requiere un tratamiento especial. Entrenar con SVM normalmente se realiza por tandas. Para grandes problemas esto significa una demanda de necesidades de memoria del ordenador. Para solventar este problema, se han ideado ciertos procedimientos. Su filosofía se basa en la descomposición, de una manera o de otra, del problema de optimización en una secuencia de otros más pequeños.

  Para grandes problemas, la fase de test también puede ser bastante exigente.

- Otra limitación importante de las máquinas de vectores de soporte es que, hasta ahora, no hay un método práctico para seleccionar la mejor función de kernel. Esto es todavía un problema sin solución.

- Las máquinas de vectores de soporte se han aplicado a una gran cantidad de de aplicaciones diversas, que van desde reconocimiento de dígitos manuscritos, el reconocimiento de objetos, identificación de personas y ecualización de canal. Los resultados indican que los clasificadores SVM ponen de manifiesto un comportamiento en general mejorado.

5.3. Agregado de SVM

Tras haber entrenado el sistema, es necesario agregar varios SVMs entrenados independientemente con un método de combinación adecuado. Ha sido demostrado que un conjunto de SVMs entrenados para la misma tarea posee mayor capacidad de generalización, son más eficaces y más robustos que un sólo SVM.

Existen dos tipos de técnicas de combinación que son los métodos lineales y los no lineales. Entre los métodos lineales, esto es, combinaciones lineales de varios SVMs, se encuentran el "Voto por mayoría", la "Ponderación basada en LSE (Least Squares Estimation)" y el "Pegado de votos". El voto por mayoría y la ponderación basada en LSE se usan habitualmente con bagging y boosting respectivamente. La idea del pegado de votos pretende aliviar problemas de requisitos de memoria para almacenar la base de datos. Por otro lado un método no lineal, es decir, combinaciones lineales de varios SVMs, incluye la combinación jerárquica de doble capa que usa otro SVM de capa superior para combinar varios SVMs de capas más bajas.

.

Parte II

Desarrollos Experimentales

CAPÍTULO 6

Bases de Datos

El conjunto de imágenes sobre el cual realizaremos nuestro estudio podrá pertenecer a tres bases de datos diferentes: un conjunto de 79 imágenes SPECT obtenidas en el Hospital Virgen de las Nieves, (Granada): un conjunto de 60 imágenes PET obtenidas en la clínica PET-Cartuja (Sevilla); y un conjunto de 219 imágenes PET obtenidas de la base de datos ADNI (Alzheimer's Disease Neuroimge Initiative) (E.E.U.U). A continuación describiremos los detalles correspondientes a cada una de ellas.

6.1. ADNI

El proyecto ADNI (Alzheimer's Disease Neuroimaging Initiative) fue iniciado en 2003 por el NIA (National Institue on Aging), NIBIB (National Institute of Biomedical Imaging and Bioengineering), el FDA (Food and Drug Administration), compañías farmacéuticas privadas y organizaciones sin ánimo de lucro, como un proyecto conjunto con financiación publico-privada que asciende a $ 60 millones de dolares. El primer objetivo de ADNI es probar si se pueden combinar las técnicas de MRI (magnetic resonance imaging), PET (positron emission tomography), otros marcadores biológicos y evaluaciones neuropsicológicas y clínicas, para medir la progresión del Alzheimer en sus primeras etapas (MCI). El descubrimiento de marcadores sensibles y específicos de las etapas mas tempranas de la enfermedad se espera que sirva de ayuda a los investigadores y médicos para desarrollar nuevos tratamientos y probar su efectividad, asi como para disminuir el tiempo y el coste de las pruebas clínicas.

El investigador principal de la iniciativa ADNI es Michael W. Weiner, M.D., VA Medical Center and University of California - San Francisco. Los pacientes de ADNI fueron seleccionados de mas de 50 lugares entre E.E.U.U y Canada. El objetivo inicial de ADNI era conseguir que 800 adultos participaran en el proyecto, con edades entre 55 y 90 años, – aproximadamente 200 sujetos normales de edad avanzada con un seguimiento previsto de 3 años, 400 personas con MCI para un seguimiento de 3 años, y 200 personas con los primeros síntomas del Alzheimer para un seguimiento de 2 años (para conocer mas detalles, visitar www.adni-info.org).

6.1.1. Protocolo de adquisición

Los escans FDG PET se tomaron todos siguiendo un protocolo estandarizado. Puesto que un gran número de centros participaron en el estudio, no todos los centros disponían de los mismos recursos materiales para efectuar los escans. Por ello, los scans fueron adquiridos según uno de los tres protocolos siguientes:

- Dinámicos: los scans de emisión dinámica consistían en 6 tomas de 5-min, adquiridas de 30 a 60 min después de la inyección intravenosa de 5.0 $\pm$0.5 mCi de ^{18}F-FDG.

- Estáticos: los estáticos consistían en una única toma de 30 min comenzando de 30 a 60 min después de la inyección (Por ejemplo los escáners Siemens PET/CT no tienen la posibilidad de tomar escans dinámicos)

- Cuantitativos: los cuantitativos consistían en un scan dinámica del doble de duración que el anterior, con 33 tomas, comenzando en el momento de la inyección y continuando 60 min. Este puede ser usado para calcular la tasa absoluta de metabolismo de glucosa, obtenida de la función de entrada del radioisotopo medida en las arterias carótidas.

La mayoría de los scans de ADNI fueron tomados siguiendo el primier protocolo. Los pacientes, a los que se solicitó que ayunaran por lo menos durante 4 h antes del scan, llacían tumbados en tranquilidad con sus ojos abiertos y la estimulación sensorial mínima.

Una vez adquiridas las imágenes, la inciativa ADNI pretende tambien minimizar en lo posible, las diferencias en los tipos de imágenes debido a la diferencia entre los scanners usados para obtenerlas. De esta manera, proporciona diferentes conjuntos de datos que se pueden descargar de sus archivos, ordenados según el tratamiento recibido. El pre-procesamiento de las imagenes ADNI esta detallado a continuación:

1. Co-registro dinámico: las imágenes PET originales de todos los sitios se descargan para el control de calidad en la Universidad de Michigan. Estas imágenes se convierten a un formato de archivo estándar. Si la imágen contiene diferentes secuencias separadas, éstas son extraídos de los archivos de imagen a efectos de registro. En la mayoría de los casos, seis secuancias de cinco minutos se adquieren entre 30 y 60 minutos después de la inyección. Cada secuencia extraída es co-registrados hasta la primera secuencia extraída del archivo de imagen original (la secuencia adquirida en los 30-35 primeros minutos después de la inyección). Las diferentes secuencias se vuelven a coregistrar a una imagen conjunto. Estas series de imágenes tienen el mismo tamaño de la imagen (por ejemplo, 128x128x63) Voxel y dimensiones (por ejemplo, 2.0x2.0x2.0 mm) y mantienen la misma orientación espacial que la imagen original PET. A esto se les denomina 'nativos'. Estos

archivos se cargan en Loni en formato DICOM. Sólo los adquiridos siguiendo el protocolo 1 ó 3, tendrá una imagen procesada de este tipo. En resumen, el 'co-registro dinámico' tiene dos principales diferencias principales con la imagen 'original' PET:

- 1) las secuencias separadas han sido co-registradas entre sí para reducir los efectos de movimiento del paciente
- 2) los archivos de la imagen están en formato DICOM.

2. Co-registrada, Promediada: Este tipo de imagen procesada conjunto se genera como un promedio de sólo 6 secuencias de cinco minutos (o las últimas 6 secuencis de los estudios cuantitativos) de la imagen de conjunto coregistrada que se ha descrito anteriormente. Esto crea una única imagen PET de 30 minutos, todavía en el espacio 'nativo'. Como en el caso anterior, sólo las exploraciones de PET adquiridas en virtud del protocolo de 1 o 3, tendrá una entrada de este tipo.

3. Co-reg, Prm, Imagen y tamaño de voxel estandarizado: Cada imagen a la que se le habían aplicado los anteriores pasos se reorientó a un estándar de 160x160x96 voxel, teniendo cada voxel un tamaño cúbico de lado 1,5 mm. Esta imagen de la red está orientada de tal manera que el eje anterior-posterior del sujeto es paralela a la línea de AC-PC. Esto se conoce como el espacio 'AC-PC' en el programa de búsqueda Loni. Esta normalización de la imagen a continuación sirve como una imagen de referencia para todas las exploraciones de PET sobre el mismo paciente. Las secuencias individuales de cada exploración PET (el estudio de referencia, así como todos los estudios posteriores (6 meses de exploración, de 12 meses de exploración, etc) son co-registrados sobre esta línea de base de referencia la imagen. Al hacer el co-registro de la imagen original en un solo paso, sólo una interpolación de los datos de la imagen es necesario, y, por tanto, la degradación de la resolución por interpolación se mantiene al mínimo, y es el mismo para todas las exploraciones. Un promedio de la imagen se genera a partir de las secuencias coregistradas 'AC-PC' de intensidad normalizada y, a continuación se utiliza una máscara específica para cada sujeto a fin de que la media de los voxels en la máscara es exactamente uno. Tanto la orientación espacial (AC-PC) como la normalización de intensidad de la imagen se toman como un punto de partida para posteriores análisis. Con una imagen de normalización de la matriz, el PET de datos de diferentes modelos

de escáner se pueden comparar más fácilmente. Cabe señalar que en estas imágenes sólo establece una nueva orientación espacial y la intensidad de la normalización de las exploraciones se ha producido. No no deformación lineal o incluso la expansión lineal de las dimensiones del cerebro se ha aplicado a las imágenes.

4. Co-reg, Prom, ESt Img y Vox, Resolución uniforme: Estas imágenes son el resultado de suavizado de las mencionadas imágenes. Cada imagen se filtra conjunto con un escáner específico de función de filtro (puede ser un filtro no isotrópico) para producir imágenes de una resolución isotrópica uniforme de 8 mm FWHM, la resolución aproximada de la resolución más baja escáneres utilizados en ADNI. Los onjuntos de imágenes de mayor resolución de los escáneres, obviamente, se han suavizado más de que las imágenes de escáneres de baja resolución. Las funciones del filtro se determinan a partir de las exploraciones de PET Hoffman fantasma que fueron adquiridos durante el proceso de certificación.

5. Los datos fueron corregidos por la radiación de atenuación y dispersión mediante la transmisión de las exploraciones de Ge-68 fuentes de rotación de la vara y reconstruido utilizando medido-corrección de atenuación y los algoritmos de reconstrucción de imágenes especificadas para cada escáner (http://www.loni.ucla.edu/ADNI/Data/ADNIData.shtml). Tras la exploración, cada imagen se examinó la posibilidad de artefactos en la Universidad de Michigan y todos datos originales y procesados fueron archivados.

En nuestro trabajo seleccionamos un conjunto de datos FDG PET de 219 participantes de ADNI, adquiridos con scanners Siemens, General Electric (GE) y Philips PET, del conjunto ofrecido por el Laboratorio de Neuroimagen ADNI LONI (Laboratory of NeuroImage, University of California, Los Angeles, http://www.loni.ucla.edu/ADNI/). En este primer estudio, los datos adquiridos con scanners Siemens HRRT and BioGraph HiRez fueron excluidos debidos a las diferencias en el patron de toma de imágenes.

6.1.2. Criterios de Etiquetado

Los criterios de elección que se siguieron para aceptar a participantes en el proyecto ADNI se basaron en una serie de entrevistas y test realizados individualmente. Los resultados de los candidatos debían cumplir ciertas

condiciones para ser admitidos en el proyecto. A continuación se detallan los criterios de selección de pacientes para cada una de las clases de interés para el estudio:

- Pacientes NC (Normal Control): La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos incluisve), un CDR de 0, no deprimido, no MCI, y sin demencia. El rango de edad de los pacientes normales sería aproximadamente coincidir con la de los sujetos AD y MCI. Por lo tanto la edad mínima para participar no debía superar los 70.
- Pacientes MCIs: La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos incluisve), debía presentar quejas por péridida de memoria, tener una pérdida objetiva de memoria medida en términos de su puntuación en el testc de Wechsler Memory Scale Logical Memory II, un CDR de 0.5, ausencia de discapacidades en otros de la función cognitiva en niveles significativos, conducta normal en las actividades de la vida cotidiana, y ausencia de demencia.
- Pacientes AD: La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos incluisve), debía presentar un CDR de 0.5 ó 1.0, y satisfacer los criterios de NINCDS/ADRDA que definen un AD probable.

Consecuentemente, los datos FDG PET se separaron en 3 clases diferentes: sujetos de control NC (Normal Controls), sujetos con afección cognitiva leve MCI (Mild Cognitive Impairment) y enfermos de Alzheimer AD (Alzheimer's Disease). En nuestro estudio contamos con imágenes de 219 sujetos diferentes, divididos en 53 AD (rango de edad: 77.2 $\pm$ 7.2 (media $\pm$ desviación estándar)), 114 MCI (rango de edad: 75.1 $\pm$ 7.4), y 52 NC (rango de edad: 76.5 $\pm$ 4.8).

6.2. Virgen de las Nieves

El rendimiento de nuestros esquemas de extracción de características junto con los clasificadores anteriormente presentados se comprueba en este trabajo sobre una base de datos que contiene 91 imágenes reales SPECT del

estudio proporcionado por el Hospital "Virgen de las Nieves" de Granada (España).

A los pacientes se les inyectó un radiofármaco emisor de rayos gamma ^{99m}Tc-ECD, adquiriendo la imagen original mediante una gamma cámara de tres cabezales Picker Prism 3000. Se tomaron 180 proyecciones con resolución angular de 2 grados. Estas imágenes de secciones del cerebro fueron reconstruidas de los datos de proyección usando el algoritmo de filtrado mediante proyección hacia atrás (FBP) en combinación con un filtro de Butterworth para eliminar ruido de alta frecuencia. Las imágenes SPECT son primeramente normalizadas usando SPM (Friston et˜al., 2007), de manera que podamos aseverar que los vóxeles de diferentes imágenes representan la misma posición anatómica subyacente en el cerebro. Este paso nos permite comparar las intensidades de los vóxeles de distintos sujetos (Salas-González et˜al., 2008).

Las imágenes SPECT fueron, a su vez, etiquetadas por los expertos del "Virgen de las Nieves" mediante 4 categorías: *normal* (NOR) para controles sanos y *AD posible* (AD1), *AD probable* (AD2) and *AD cierto* (AD3) para distinguir entre distintos niveles de la presencia del patrón de AD. En total, la base de datos contiene 41 NOR, 27 AD1, 19 AD2 y 4 AD3.

6.3. PET Cartuja

Las técnicas de diagnóstico PET permiten la diferenciación entre tejido enfermo y sano. Consiste en el registro de imágenes que presentan en 3 dimensiones la distribución orgánica de la molécula más utilizada, la FDG (solución de glucosa marcada con flúor 18). En el cerebro, la distribución de FDG está directamente relacionada con la actividad neuronal y, en el miocardio refleja la actividad cardíaca. Las cámaras PET recogen las señales (radiación) y reproducen en imágenes los procesos detectados.

El protocolo de adquisición seguido en PET-Cartuja se detalla a continuación: La adquisición de las imágenes no debe comenzar hasta pasados 30 minutos desde la administración de la ^{18}FDG. En este tiempo el paciente permanece en reposo en una habitación en silencio y con iluminación tenue para que el contraste se distribuya adecuadamente por todo el organismo. Se

recomienda fijar un tiempo estándar, por ejemplo, 30 minutos, para que los estudios de distintos pacientes a los de distintos controles sean comparables. El paciente se posiciona en decúbito supino. Debe dedicarse un sistema dedicado para apoyar confortablemente la cabeza, fijándola mediante cintas para evitar movimientos involuntarios. El paciente se situa acostado en la camilla de la cámara y se desplazará progresivamente por el centro del anillo de la Cámara PET durante un tiempo aproximado de 30 minutos en un estudio de PET cerebral.

Durante este tiempo la cámara PET recoge las señales emitidas por el contraste en todo el cuerpo. Después, un ordenador recoge las señales emitidas y las convierte en imágenes funcionales tridimensionales en los tres planos del espacio (axial, coronal y sagital) y una imagen volumétrica del organismo en tres dimensiones. Los positrones emitidos por el radiofármaco colisionan con los electrones (con carga negativa) de los átomos que componen las moléculas tisulares. La interacción positrón-electrón origina la aparición de un par de fotones con el aniquilamiento de las masas del positrón y electrón. Estos dos fotones presentan una energía de $512KeV$ cada uno, y se desplazan en la misma dirección y en sentidos opuestos, excitando de forma simultánea 2 detectores de la cámara PET que se encuentran en coincidencia en un ángulo de 180°. Esta detección permite "por coincidencia" la reconstrucción tomográfica tridimensional del organismo que representa la distribución tisular del radiofármaco.

En este estudio contamos con imágenes PET de 60 pacientes diferentes, divididos en 47 AD y 13 NC, lo que supone una prevalencia de la clase AD que ha de tenerse en cuenta en análisis posteriores.

CAPÍTULO 7

Pre-Procesado de Imágenes

El valor de una técnica de asistencia al diagnóstico por computador depende de manera fundamental en una tecnica efectiva de adquisición, asi como de una reconstrucción y registro apropiados. Una vez introducidos todos los conceptos y herramientas necesarios para construir el sistema de diagnósitico asistido, en esta sección presentaremos los pasos previos de preprocesamiento que es necesario aplicar a las imágenes adquiridas antes de definir los métodos de clasificación. Los dos primeros pasos, en concreto la adquisición de imágenes 7.1 y la reconstrucción de ellas 7.2, serán partes sobre las que no hemos tenido ningún control en nuestro trabajo, ya que los datos manejados habrán sido sometidos a estos pasos previos en los centros médicos. Puesto que un número elevado de centros diferentes tomarán parte en los datos que manejaremos, sobre todo debido al uso de la base de datos ADNI, tomaremos el ejemplo de las imágenes SPECT de Virgen de las Nieves para ilustrar una metodología concreta para llevar a cabo estos dos pasos de preprocesamiento. Los siguientes pasos sí estarán bajo nuestro

control, y serán comunes a todas las bases de datos, por lo que se describirán de forma genérica.

7.1. Adquisición de Imágenes

En el proceso de adquisión de imágenes SPECT el paciente se posiciona cómodamente en una camilla con la cabeza "inmobilizada". El detector debe posicionarse tan próximo al cerebro del paciente como sea posible, preferiblemente con un radio de rotación de 14 cm o menos desde la superficie del detector de colisiones al centro del cerebro del paciente. Al paciente se le inyecta un fármaco emisor de rayos gamma ^{99m}Tc-ECD y el scan SPECT se adquiere gracias a una cámara de tres cabezales Picker Prism 3000. Se toman un total de 180 proyecciones por cada paciente con una resolución angular de 2 grados. Finalmente, las imágenes de las secciones cruzadas del cerebro se consiguen mediante la reconstrucción a partir de las proyecciones usando el algoritmo de retroproyección filtrada (FBP) descrito a continuación en combinación con un filtro Butterworth para la eliminación de ruido.

7.2. Reconstrucción de Imágenes

Las imágenes de sección eficaz del cerebro pueden reconstruirse a partir de los datos de proyección (Lange and Carson, 1984; Vardi et˜al., 1985; Hudson and Larkin, 1994; Bruyant, 2002; Chornoboy et˜al., 1990). En condiciones ideales, las proyecciones son conjuntos de medidas de los valores integrados de algunos parámetros del objeto. Si el objeto se representa por una función bidimensional $f(x, y)$ y cada integral de línea por los parámetros (θ, t) la integral de línea se define como:

$$P_\theta(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - t)dxdy, \qquad (7.1)$$

donde $P_\theta(t)$ se conoce como la transformada Radon de la función $f(x, y)$.

La clave en imagen tomográfica es el *Teorema de Fourier de Rodajas* el cual relaciona las medidas de los datos de proyección con la transformada de Fourier de la sección eficaz del objeto. El teorema establece lo siguiente:

Teorema: La transformada de Fourier $S_\theta(w)$ de la proyección paralela $P_\theta(t)$ de una imagen $f(x, y)$ tomada con un ángulo θ y definida del siguiente modo:

$$S_\theta(w) = \int_{-\infty}^{+\infty} P_\theta(t) \exp(-j2\pi wt)dt, \tag{7.2}$$

proporciona una rebanada de la transformada de Fourier bidimensional:

$$F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) \exp(-j2\pi(ux + vy))dxdy, \tag{7.3}$$

que se encuentra a un ángulo θ del eje u, es decir,

$$S_\theta(w) = F(u = w\cos\theta, v = w\sin\theta). \tag{7.4}$$

El resultado de arriba es la esencia de la tomografía de rayo directo e indica que teniendo proyecciones de un objeto con ángulos θ_1, θ_2, ..., θ_k y tomando su transformada de Fourier, los valores de $F(u, v)$ pueden determinarse en líneas radiales. En ese caso, es claro que la función $F(u, v)$ se conoce sólo a lo largo de un número finito de líneas radiales de modo que uno debe interpolar dichos puntos radiales a los puntos de la rejilla cuadrada.

Los datos de proyección usados en este estudio se reconstruyen usando el algoritmo de retroproyección filtrada (FBP) que se deriva fácilmente del teorema de Fourier de rodajas. Una imagen de la sección cruzada $f(x, y)$ de un objeto se puede recuperar por:

$$f(x, y) = \int_0^{\pi} Q_\theta(x\cos\theta + y\sin\theta)d\theta, \tag{7.5}$$

donde

$$Q_\theta(t) = \int_{-\infty}^{+\infty} S_\theta(w)|w| \exp(j2\pi wt)dw. \tag{7.6}$$

El algoritmo FBP consiste entonces en dos pasos: la parte de filtrado, la cual puede verse como una simple ponderación de cada proyección en el dominio frecuencial, y la parte de proyección hacia atrás.

Un gran inconveniente de FBP es que se amplifica de forma indeseada el ruido de altas frecuencias impactando sobre la calidad de la imagen. Estos efectos se producen por la de multiplicación de $S_\theta(w)$ por $|w|$ en la ecuación

7.6. Para atenuar el ruido de altas frecuencias amplificado durante la reconstrucción FBP se han propuesto diversas funciones tipo ventana. De este modo, el método de reconstrucción descrito por las ecuaciones 7.5 y 7.6 se redefine normalmente aplicando una ventana en frecuencia con valores cercanos a cero cuando la frecuencia tiende a π. Entre las ventanas más comunes para la reconstrucicón FBP se encuentran:

- sinc (filtro de Shepp-Logan)
- coseno
- Hamming
- Hanning

Sin embargo, incluso cuando el ruido de reconstrucción se mantiene bajo usando la aproximación de FBP con control de ruido, se necesita filtrar el ruido capturado por el sistema de adquisición para mejorar la calidad de las imágenes reconstruidas. Además, la etapa de preprocesado de la mayoría de sistemas de preprocesado de imágenes SPECT a menudo incorpora prefiltrado, reconstrucción y postfiltrado para minimizar el ruido captado por la cámara gamma así como el ruido amplificado por la reconstrucción FBP.

7.3. Registro de Imágenes

La complejidad de las estructuras cerebrales y las diferencias entre cerebros de diferentes sujetos hace que la normalización de las imágenes con respecto a una plantilla común sea necesaria. Este paso nos permite comparar las intensidades de los voxels de las imágenes cerebrales de pacientes diferentes. Las imágenes son inicialmente normalizadas espacialmente usando la herramienta SPM (Statistical Parametric Mapping) (Friston et˜al., 2007)(http://www.fil.ion.ucl.ac.uk/spm/software/spm5) para asegurarnos de que los voxels de distintas imágenes se refieren a la misma posición anatómica del cerebro. El método de normalización asume un modelo genérico afín con 12 parámetros (Woods, 2000) y una función de coste la cual presenta un valor extremo cuando la plantilla y la imagen se corresponden una con la

otra (Ramírez et˜al., 2008). La función objetivo que se ha de optimizar es la diferencia cuadrática media entre ambas imágenes, la fuente y la plantilla.

$$\text{CF} = \sum_i (f(\mathbf{M}\mathbf{x}_i) - g(\mathbf{x}_i))^2, \tag{7.7}$$

donde f denota la imagen original y g la plantilla. Para cada voxel $\mathbf{x} = (x_1, x_2, x_3)$ de una imagen, la transformación afín a las coordenadas $\mathbf{y} = (y_1, y_2, y_3)$ se expresa mediante la multiplicación matricial $\mathbf{y} = \mathbf{M}\mathbf{x}$.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}. \tag{7.8}$$

Tras la normalización afín, la imagen resultante se registra usando un modelo de transformación no rígido más complejo. Las deformaciones no lineales se ajustan a una plantilla de Montreal Neurological Imaging (MNI), parametrizándose por una combinación lineal de componentes de las más bajas frecuencias de las bases de la transformada coseno tridimensional (Ashburner and Friston, 1999). Pueden adoptarse modelos de transformación afín y normalización espacial no ineal de convergencia más rápida, como el descrito en (Salas-González et˜al., 2008).

La Figura 7.1 muestra un ejemplo de la operación del proceso de normalización en imágenes SPECT. La columna izquierda muestra una imagen fuente arbitraria de la base de datos, la columna central muestra la plantilla usada en el registro de la imagen, y finalmente la imagen normalizada correspondiente en la columna derecha. Tras usar un modelo de deformación pequeña, y una regularización de la energía de curvatura del campo de desplacamiento, se observa claramente cómo la imagen transformada se ajusta a la forma de la plantilla.

Tras la normalización espacial, una vez garantizada la correspondencia espacial de los voxeles entre distintas imágenes, se obtiene una representación de cada individuo de 95×69×79 voxels tanto para imágenes SPECT como para PET. Cada voxel representa un volumen cerebral de 2.18×2.18×3.56 mm^3. Este paso es esencial a la hora de hacer posible comparaciones entre

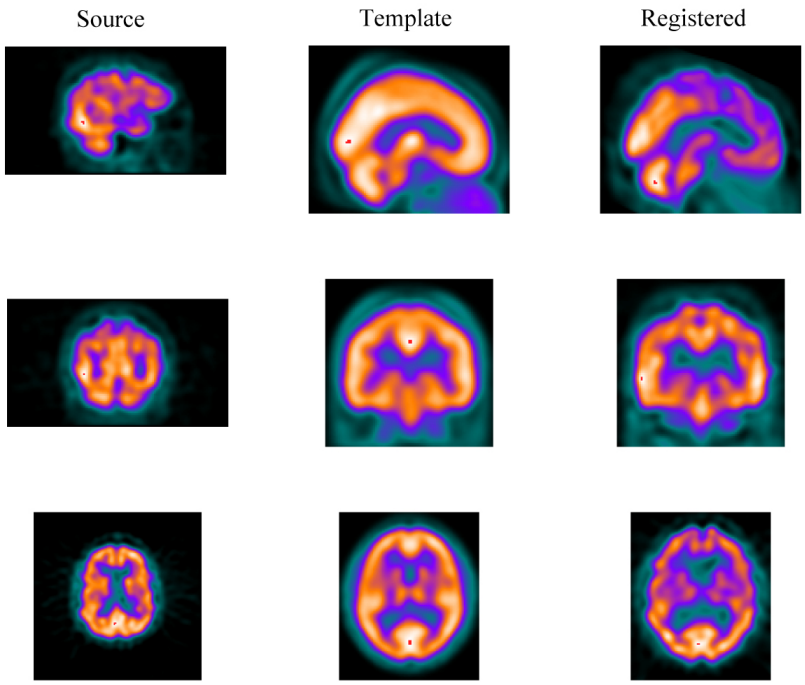

Figura 7.1: Tres imágenes SPECT. *Columna izquierda*: Imagen original. *Columna central*: Plantilla. *Columna derecha*: Imagen transformada tras el proceso de registro.

imágenes, y nuestros resultados dependerán en gran medida de éste paso previo. El reto de la normalización espacial supone conseguir minimizar los efectos que producen las diferencias en la imágen debidas a características indiviuales de cada sujeto, sin alterar las diferencias debidas a efectos de la enfermedad, lo que en la práctica es díficil de conseguir completamente. La dificultad aumenta en gran medida debido al hecho de tratar con información funcional, que hace que los límites espaciales de las diferentes regiones cerebrales sean difusos.

7.4. Normalización de intensidad

Después de la normalización espacial, se requiere una normalización en intensidad para poder comparar entre imágenes de diferentes sujetos. La comparación directa entre la información de la intensidad de los voxels, incluso entre diferentes adquisiciones de un mismo sujeto, no es posible sin ésta normalización de las intensidades.

Llevando a cabo un proceso parecido a (Saxena et˜al., 1998), si el nivel de intensidad se normaliza a la intensidad máxima, la cual se calcula individualmente para cada volumen calculando la media del 3 % de los voxels de mayor intensidad, las imágenes pueden estar expuestas a saturación. Como se puede observar en la figura 7.2, los patrones de intensidad que presentan las imágenes pueden ser muy variados, aunque todos presentan una acumulación de intesidad en los primeros niveles de intesidad. Esta acumulación de intesidad se corresponde con las regiones de fuera del cerebro, que suponen una gran cantidad de información irrelevante. Si se toma el 3 % de todos los voxels, se tendrá en cuenta mucha de esta información lo que producirá que se normalice a un máximo de intensidad relativamente bajo, y consecuentemente las imágenes saturarán.

Nuestra propuesta parte de una base similar, normalizando la intensidad de las imágenes a un valor máximo I_{max}, pero difiriendo en su obtención. Esta se lleva a cabo promediando el 0.1 % de los voxeles de mayor intensidad que superan un umbral. El umbral se prefijó ajustandose manualmente a un valor optimo, obtenido como el valor de intensidad del décimo bin en un histograma de intensidad de 50 bins, asegurandonos que la información irrelevante de baja intensidad contenida fuera del cerebro fuese deshechada, y previniendo la saturación.

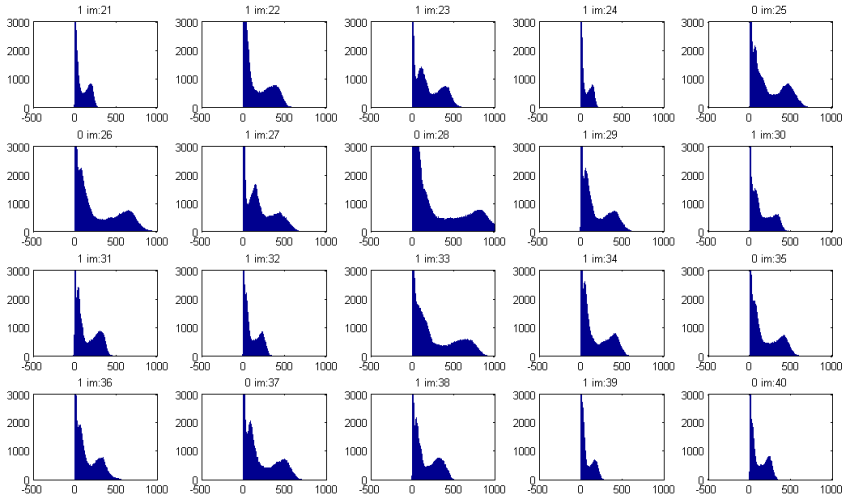

Figura 7.2: 20 histogramas de intensidad de diferentes imágenes de la base de datos SPECT

Parte III

Análisis en Componentes de Imágenes Funcionales

CAPÍTULO 8

Componentes

Esta sección muestra una novedosa técnica de extracción de características de ayuda al diagnóstico que combina técnicas de reducción de características basadas en métodos de filtro de tipo máscara 4.3.1 y de envoltura 4.3.2 con el método estadístico de pegado de votos (Breiman, 1999) para la agregación de clasificadores SVM 5.3 y su aplicación a las componentes relevantes de la imagen. La propuesta se define como un esquema híbrido dado que aplica un análisis multivariado usando ROIs pero, a la vez, considera una reducción de características y selección de componentes del espacio de entrada que se acerca a la aproximación univariada (la decisión, que en SPM se obtiene a nivel de vóxel, en nuestro caso se toma sobre un conjunto de vóxeles o componentes a la cual aplicamos el clasificador). Siguiendo la metodología de SPM, la inferencia se evalúa relacionando análisis previos en cada componente.

8.1. Método de Componentes

La mayor parte de los procesos biológicos, como por ejemplo la actividad cerebral, exhiben localidad: las características que contienen información de regiones anatómicas que están "próximas", tienen probabilidad de estar altamente correlacionadas. Además, en nuestro caso, el diagnóstico efectivo de la enfermedad del Alzheimer se basa en el hallazgo de pacas de amiloide y enredos de neurofibrillas en determinados áreas corticales en un suficiente número en la autopsia, lo que implica una localidad de las áreas afectadas. Una técnica basada en estadística univariada, por ejemplo la que usa información a nivel de vóxel para la clasificación, es insuficiente para recuperar información fundamental que reside en la influencia entre vóxeles. De hecho se sabe que sólo los vóxeles adyacentes en ciertas regiones cerebrales serían relevantes en la distinción de los pacientes afectados por la enfermedad del Alzheimer.

Por lo tanto, el uso de técnicas estadísticas multivariadas aplicadas a imágenes funcionales para la ayuda al diagnóstico de la enfermedada del Alzheimer está plenamente justificado para superar las limitaciones impuestas por las aproximaciones univariadas ampliamente usadas, como SPM (Friston et˜al., 2007; Ziolko et˜al., 2006). No todas las técnicas multivariadas serán adecuadas para dar cuenta de las relaciones locales existentes entre vóxeles. Por ejemplo, VAF 2 considera todos los vóxels en un solo vector teniéndose en cuenta las relaciones entre ellos, aunque se pierde la información local, ya que interesan solo las relaciones entre un grupo pequeño de voxels, que corresponde con la región cerebral afectada. Ademásque presenta otro tipo de limitaciones como la necesidad de un número de muestras elevado. SVM se aplica como ejemplo de análisis multivariado, pero aplicado directamente presenta el problema de la maldición de la alta dimensión, al igual que la citada técnica estadística para inferencia Mancova, debido al hecho de que la imagen representa un gran cantidad de información y la mayor parte de estudios en neuroimagen disponen de un número de muestras pequeño (generalmente <100) (Ishii et˜al., 2006).

El enfoque que proponemos en esta sección está basado en la descomposición de una imagen funcional, por ejemplo una imagen SPECT, en un conjunto de subimagenes o *componentes*. La principal motivación para hacer esta factorización es la búsqueda de la regiones más relevantes para la

clasificación. Cada componente corresponderá a una región cerebral, y se hará un análisis individual de cada componente a través de un SVM. Esto permitirá localizar las regiones interesantes e ignorar as áreas cerebrales irrelevantes para la clasificación. Al mismo tiempo, el problema del tamaño muestral pequeño también se solventa por medio de este esquema implícito de reducción de la dimensión de las características: para estudiar independientemente las componentes, se construirá un vector de características con ellas, obteniéndose uno de menor dimensión. Estas regiones pueden estar separadas en el espacio, por lo que es necesario aplicar finalmente un agregado de SVMs para obtener la decisión final sobre el sujeto de estudio.

8.2. Extracción de características basada en Factorización

Una imagen funcional es una representación 3D del cerebro que reconstruye en un volumen $V \subset \mathbb{R}^3$ algún proceso que tiene lugar en el cerebro. Dependiendo de la técnica de imagen empleada, éste puede ser información del riego sanguíneo, actividad metabólica de la glucosa, etc. La factorización del volumen consiste en la división de toda la imagen cerebral en un conjunto de subvolumenes ó componentes, para realizar la tarea de clasificación sobre cada componente. De manera explícita, describamos una imagen funcional de un sujeto como una función de 3 variables $\mathbf{I}(x, y, z)$, que contiene la intensidad registrada en cada vóxel con coordenadas $(x, y, z) \in V$. Debido a las limitaciones técnicas, solo se podrá medir un conjunto de valores de esta función continua, ya que las posiciones de los vóxeles (x, y, z) forman una red cúbica y la distribución de intensidad dentro del volumen V tomará valores discretos como las imágenes muestreadas en 2D. Por lo tanto, la información real registrada de la imagen funcional del cerebro será representada por una matriz $\mathbf{I}$ de tamaño $X \times Y \times Z$, donde el tamaño de la matriz da cuenta de la longitud de los ejes que delimitan el volúmen cúbico V que encierra la información cerebral. El elemento de matriz I_{ijk} será la intensidad medida en el vóxel con coordenadas (x_i, y_j, z_k), es decir:

$$I_{ijk} = I(x_i, y_j, z_k), \quad \begin{matrix} i = 1, .., X \\ j = 1, ..., Y \\ k = 1, ..., Z \end{matrix} \tag{8.1}$$

Considérese el conjunto

$$C = \{(x_a, y_b, z_c) : (x_a, y_b, z_c) \in C \subset V\} \tag{8.2}$$

que define un subvolumen de V. Notese que la definición del subvolumen C no hace referencia a la forma que ha de tener este, a pesar de que V tiene una fórma cúbica. Un ejemplo de C sería una esfera con un radio suficientemente pequeño para quedar completamente contenida en el interior de V. En la práctica, puesto que V esta muestreado en vóxels, C estará formado por un conjunto discreto de vóxels, y contendrá un numero finito de elementos F.

Podemos considerar dividir el volumen V en s subvolumenes $C_1, C_2, ..., C_s$, de manera que todo el volumen V quede cubierto. La imagen cerebral completa $\mathbf{I}(V)$ queda subdividida en el mismo número de subconjuntos ó componentes $\mathbf{I}(C_1), \mathbf{I}(C_2)$, ..., $\mathbf{I}(C_s)$, donde una *componente* estará constituida por un conjunto de valores de intensidad de manera que:

$$\mathbf{I}(V) = \bigcup_{m=1}^{s} \mathbf{I}(C_m) - \bigcap_{m=1}^{s} \mathbf{I}(C_m) \tag{8.3}$$

donde el segundo término de la parte derecha de la igualdad elimina la redundancia del solapamiento de las componentes, por ejemplo, las componentes que se solapan y cubren parcialmente la misma región cerebral[1]. Cada componente $\mathbf{I}(C_m)$ selecciona una región del cerebro mediante un conjunto de vóxeles (see Fig. 8.1). Las intensidades en esos vóxeles son concatenadas en un vector de características $\mathbf{x} = (x_1, ..., x_F)$, cuyas coordenadas x_i vienen dadas por

$$x_i = I_{abc} = I(x_a, y_b, z_c), \ (x_a, y_b, z_c) \in C_m \tag{8.4}$$

[1]El motivo de considerar este caso es el considerar la búsqueda "fina" de localizaciones de componentes para poder adaptarlas la las regiones cerebrales más relevantes para la clasificación, discutido en la siguiente sección.

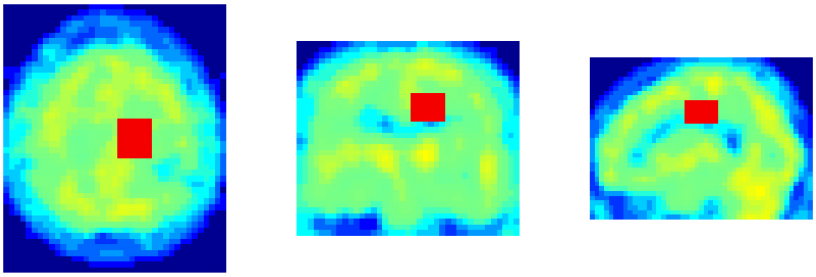

Figura 8.1: Cortes sagital, coronal y transversal de una imagen SPECT cerebral con la componente de la imagen remarcada

y donde F es el número de vóxeles en esa componente. Cada vector $\mathbf{x} \in \mathbb{R}^F$ constará de una etiqueta con $y \in \pm 1$. Estos vectores etiquetados se usan como vectores de características para la construcción de un clasificador SVM. Exisirán tantos vectores por cada imagen cerebral como subdivisiones en componentes, todos ellos compartiendo la misma etiqueta. Sin embargo, la tarea de clasificación se llevará a cabo considerando cada componente individualmente, obteniéndose un numero s de caterorizaciones alternativas de una misma imagen. El último paso supondrá el agregado del conjunto de decisiones SVM para obtener una decisión final colectiva.

8.3. Agregación del Conjunto de Decisiones basadas en SVM

En nuestra base de datos contaremos con un conjunto de n imagenes funcionales. Cada imagen funcional será factorizada en s componentes, construyendose con los datos originales un conjunto de vectores de características:

$$\mathbf{x}_i^{(m)} \in \mathbb{R}^F, \quad \begin{matrix} i = 1, ..., n \\ m = 1, ..., s \end{matrix} \tag{8.5}$$

para que un clasificador SVM categorice cada una de estas componentes en cada uno de los pacientes. Al igual que SPM requiere el uso de la Teoría

de Campos Aleatorios (Friston et˜al., 2007) para realizar la inferencia estadística, nuestro modelo está basado en la técnica de "pegado" de votos para alcanzar la decisión final sobre el individuo (Breiman, 1999). Nuestra aproximación para la agregación promedio de SVMs requiere el uso de dos etapas dependiendo si se emplea de selección de componentes o no:

8.3.1. Voto por mayoría

En este esquema se usan todas las componentes para la clasificación. El clasificador SVM se entrena para todos los pacientes $\mathbf{x}_i^{(m)} i = 1, ..., N-1$ excepto uno, usando éste para test. Siguiendo la notación de (4.2), se obtendrá una categorización para él:

$$f\left(\mathbf{x}_{test}^{(m)}\right) \equiv z^{(m)},\ z^{(m)} \in \{\pm 1\} \tag{8.6}$$

repitiéndose en cada componente, es decir de $m = 1$ a $m = s$, proporcionando un conjunto de s decisiones o votos para el paciente. El conjunto de decisiones SVM nos servirán para definir una decisión global basándonos en la técnica de pegado de votos propuesta en (Breiman, 1999), que consiste en la suma no pesada de los s votos que proyecta cada componente sobre un paciente concreto. Explícitamente, esta función $\mathcal{F}$ de decisión colectiva, que denominaremos *voto por mayoría*, que se define como:

$$\mathcal{F}(z) = \sum_{m=1}^{s} z^{(m)} \tag{8.7}$$

clasificando el paciente como *normal* si $\mathcal{F}(z) > 0$, y como AD si $\mathcal{F}(z) < 0$. Este proceso se itera n veces, hasta que cada paciente de la base de datos es usado una vez como test (técnica de validación dejar un fuera).

El carácter democrático de este método de agregación se plasma en el hecho de que el paciente será clasificado como perteneciente a una clase si la *mayoría* de las componentes de ese paciente son clasificadas como pertenecientes a esa clase, a pesar de que una cantidad destacable de componentes se clasifiquen erroneamente.

Este carácter democrático ofrece una metodología subyaciente muy interesante de extracción de características relevantes para clasificación: Las regiones del cerebro irrelevantes para la detección de la enfermedad serán clasificadas de manera aleatoria en una u otra clase, mientras que las relevantes categorizarán el paciente de la misma manera. Si el número de componentes es sufientemente grande, los votos emitidos aleatoriamente tenderán a compensarse, haciendo que la suma total de los votos de regiones irrelevantes sea nula. De esta manera, unos pocos votos emitidos de manera no aletoria harán inclinar la decisión final hacia una clase concreta, haciendo que se seleccionen de forma indirecta los votos de las regiones de interés.

De esta manera, el método de voto por mayoría destaca por su sencillez y efectividad, con la interesante propiedad añadida de ser muy robusto, ya que debido a lo anteriormente expuesto, los errores tienden a compensarse y es débilmente afectado por la modificación en cualquiera de los parámetros.

8.3.2. Voto por Relevancia

En este esquema permitimos que solo las componentes más relevantes proporcionen su voto. La relevancia se define en cada componente en función del rendimiento que proporciona el clasificador sobre esa componente, siendo las más relevantes aquellas en las que el proceso de clasificación sea más eficaz.

Para ordenar las componentes de una imagen según un criterio de relevancia, será necesario tener una estimación del rendimiento del clasificador sobre ellas. Por lo tanto, es necesario el uso de un filtro de tipo envoltura, que permitirá seleccionar las componentes más relevantes. Según vimos en la sección 4.3.2, este tipo de filtros utiliza la clasificación internamente como medio de seleccionar características, estimando la precisión del clasificador sobre un conjunto de características, para después descartar aquellas que empobrecen las tasas de clasificación. Mientras, un conjunto de elementos se mantienen fuera del proceso, para ser usados como conjunto de test.

De manera específica, seleccionamos un conjunto de l pacientes de la base de datos para la extracción de características, mientras que dejamos fuera $N - l$ para el test. Cada imagen de entrenamiento se factoriza y las componentes resultantes $\mathbf{x}^{(m)}$, se utilizan para obtener s clasificadores SVM:

$$f\left(\mathbf{x}^{(m)}\right) \equiv z^{(m)},\ m = 1, ..., s \tag{8.8}$$

lo que proporciona un conjunto de s decisiones SVM sobre cada paciente. Mediante la estrategia de validación cruzada de dejar uno fuera, se obtiene un conjunto de decisiones para cada paciente.

Comparando cada valor de $z(m)$ con la etiqueta original, es posible determinar si la componente categoriza el paciente como PV, PF, NV o NF. Estudiando esta categorización en cada paciente se conocerá el número de PV, PF, NV y NF que el clasificador produce en esa componente. Con esta información, se podrá calcular la Precisión, Sensibilidad, Especificidad,...etc, de cada una de las componentes. Tanto Precisión, Sensibilidad, como Especificidad proporcionan valores numéricos que se pueden usar como criterio para determinar la relevancia de una componente. Designando genericamente como A a este valor numérico, que puede referirse a cualquiera de las tres funciones anteriores o incluso a otra diferente, se podrá asignar un valor $A^{(m)},\ m = 1, 2, ..., s$ a cada componente.

Construiremos con estos valores el conjunto de componentes relevantes, que serán todas aquellas cuyo valor de A sea suficientemente alto. El criterio para determinar si A es o no alto será compararlo con un valor estándar T considerado como alto. Así, todas las componentes cuyo valor de A sea superior a T serán consideradas relevantes, y pertenecerán al conjunto de componentes relevantes:

$$G_T = \{m : A^{(m)} > T\} \tag{8.9}$$

Establecido un criterio de relevancia, se define la función de clasificación $\mathcal{T}$ (que denominaremos *voto por relevancia*) sobre un paciente, como la suma no pesada de las componentes relevantes, que equivale a la suma de aquellas componentes cuya Precisión, Sensibilidad, Especificidad,... sea mayor que un cierto umbral prejidado. Matemáticamente, el número limitado de votos vendrá determinado por:

$$\mathcal{T}(z) = \sum_{m \subset G_T} z^{(m)} \tag{8.10}$$

Limitamos el número de votos a $s' < s$, mediante un umbral T que prefijaremos para determinar la relevancia de las componentes. Así, sólo un subconjunto de componentes G_T con valores $A^{(m)}$ mayores de T se tienen en cuenta para la votación. Como para el caso de la función de decisión $\mathcal{F}$ de voto por mayoría, el valor del signo de la salida en $\mathcal{T}$ definirá la pertenencia de clases y dependerá del valor de T elegido.

La selección de características generalmente conduce a la obtención de clasificadores con un alto poder de generalización, y mediante la extracción de estas características relevantes podemos localizar los atributos que son responsables de las diferencias entre las clases bajo estudio. Por lo tanto, el voto por relevancia basado en el método de envoltura puede proporcionar una solución más precisa (Kohavi and John, 1997) que otros métodos, pero en general la complejidad computacional aumenta de manera sensible. Este aumento sugiere el uso del método de K-pliegues como validación, testando el rendimiento de la función de decisión (8.10) sobre el conjunto de los restantes $N - l$ sujetos, e iterando K veces el proceso.

Es interesante destacar que el subconjunto G_T dado por (8.9), define una máscara que selecciona las regiones de mejor rendimiento en la clasificación, que corresponderán con las regiones relevantes para el diagnóstico del Alzheimer. Esta máscara se puede asumir independiente del conjunto muestral, a pesar de que el patrón de Alzheimer es variable, y que caracteriza a la propia enfermedad. Debido a la independencia de esta máscara con la base de datos, esta puede ser computada "off-line", reduciendo el coste computacional de la aproximación.

Por otro lado, es interesante resaltar que los médicos suelen usar solo algunas zonas para la evaluación visual de los pacientes. Existe un gran número de trabajos donde se detallan que zonas son las más adecuadas para esta tarea, a pesar de que no hay una respuesta final a la pregunta cuales son las zonas más discriminantes a este respecto. Por ejemplo se ha mostrado que ciertas regiones que presentan un "poder predictivo" elevado existe un número considerable de pacientes que no muestran esos signos de actividad (Nitrini et˜al., 2000). En particular, existen tres zonas claras mencionadas en la literatura para el diagnóstico de la AD como son (Goethals et˜al., 2002) la región temporo-parietal, el cingulado posterior y el lóbulo temporal. Estas regiones se denominan ROIs y puede apreciarse en la figura 8.2, que dado un umbral T, el método propuesto de voto por relevancia define las mismas.

8.4. Experimientos

Se realizaron dos experimentos de agregado. Primero, las imágenes cerebrales fueron factorizadas usando una división cúbica extrayendo un número de componentes M, con un número de vóxeles fijo en cada componente (el valor F en (8.4)), ó con un número fijo de componentes permitiendo que el número de componentes variase. La base completa se usó para el entrenamiento y test de SVMs mediante un método de validación basado en la estrategia de validación cruzada de k-pliegues, con 11 pliegues, para acelerar el proceso de aprendizaje.

El segundo experimento fue diseñado específicamente para el método del voto por relevancia. DEspués del mismo proceso de factorización sin fijar el número de vóxeles. Usamos un subconjunto de $l = 53$ pacientes para entrenar y testar SVMs por medio de la estrategia de de validación dejar uno fuera. Con ellos se efectuó una búsqueda exhaustiva en todo el cerebro, para la detección de las ROIS más importantes. De esta forma construimos la máscara ROI, para después usarla para testar el conjunto o agregado de SVMs en los restantes pacientes. El resultados del test en éstos se ´"agregó" usando un valor de $\mathcal{T}$ de la función de decisión para distintos valores del umbral T. Este proceso fue embebido en 3-pliegues para obtener una mejor estimación de los parámetros del rendimiento asegurando de esta forma que estos parámetros no dependieran del conjunto retenido dado que promediamos sobre el conjunto de estos tres pasos.

8.5. Resultados

Los resultados de este procedimiento se muestran en la tabla (8.1), donde el agregado se obtuvo mediante la técnica de voto en mayoría.

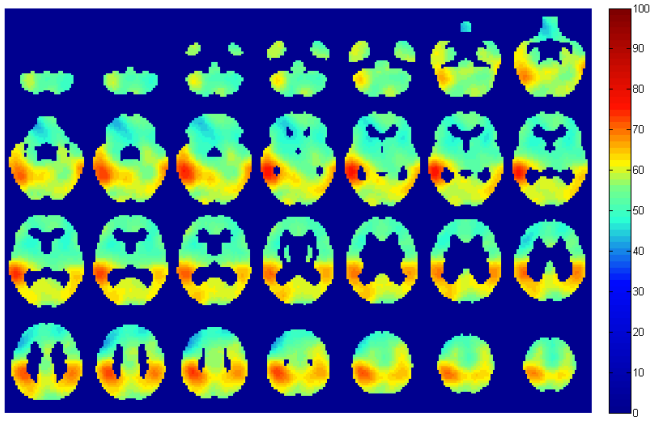

Figura 8.2: Mapa de las ROIs de un conjunto de imágenes SPECT obtenidas mediante el valor de A_i

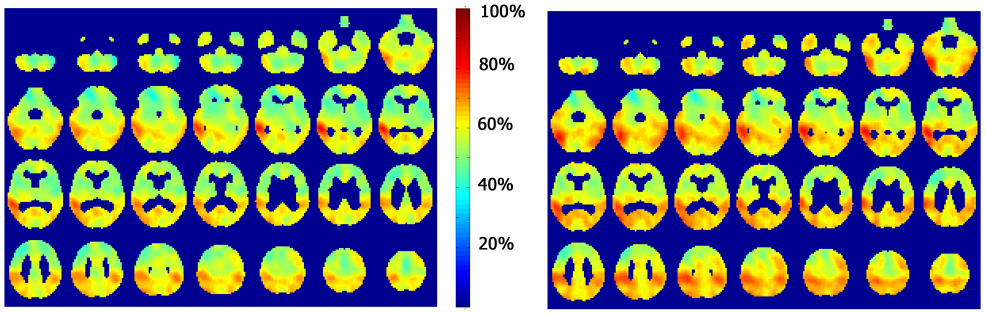

Figura 8.3: Mapa de las regiones de interés según su precisión considerando solo pacientes ATD-1 (izquierda) o considerando solo pacientes ATD-2 y ATD-3 (derecha)

Reducción	Parámetro (%)	Número de Vóxel			# Componente			VAF
		15	20	25	20	100	750	
Factor $1/4^3$	Precisión	86.08	86.08	82.28	68.35	74.68	72.15	74.68
	Especificidad	87.80	87.80	82.93	73.17	82.93	80.49	82.93
	Sensibilidad	84.24	84.24	81.58	63.16	65.79	63.16	65.79
Factor $1/6^3$	Precisión	83.54	86.08	86.08	70.89	77.22	73.42	72.15
	Especificidad	87.80	90.24	90.24	82.93	87.80	82.93	78.05
	Sensibilidad	78.95	81.58	81.58	57.89	65.79	63.16	65.79
Factor $1/8^3$	Precisión	81.01	79.75	83.54	72.15	69.62	75.95	64.56
	Especificidad	85.37	85.37	87.80	85.37	82.93	85.37	70.73
	Sensibilidad	76.32	73.68	78.95	57.89	55.26	65.79	57.89

Tabla 8.1: Medidas estadísticas del rendimiento de la prueba por voto en mayoría, con un número fijo de vóxeles contenidos en cada componente ó con un número fijo de componentes. VAF (Stoeckel et˜al., 2001) se muestra como referencia.

Reducción	Parámetro (%)	Tamaño Componente				VAF
		$1/3L$	$1/4L$	$1/5L$	$1/6L$	
Factor $1/4^3$	Precisión	86.18	84.81	86.13	83.48	74.68
	Especificidad	90.28	90.48	93.75	93.17	82.93
	Sensibilidad	80.68	81.58	80.95	63.16	65.79
Factor $1/6^3$	Precisión	89.03	86.13	84.90	82.15	72.15
	Especificidad	87.73	95.14	91.67	76.46	78.05
	Sensibilidad	84.54	74.29	79.32	73.89	65.79
Factor $1/8^3$	Precisión	86.09	80.82	74.34	65.81	64.56
	Especificidad	89.68	87.92	78.93	68.68	70.73
	Sensibilidad	81.67	77.27	70.04	63.68	57.89

Tabla 8.2: Medidas estadísticas para el voto por relevancia, véase el texto para una mayor descripción.

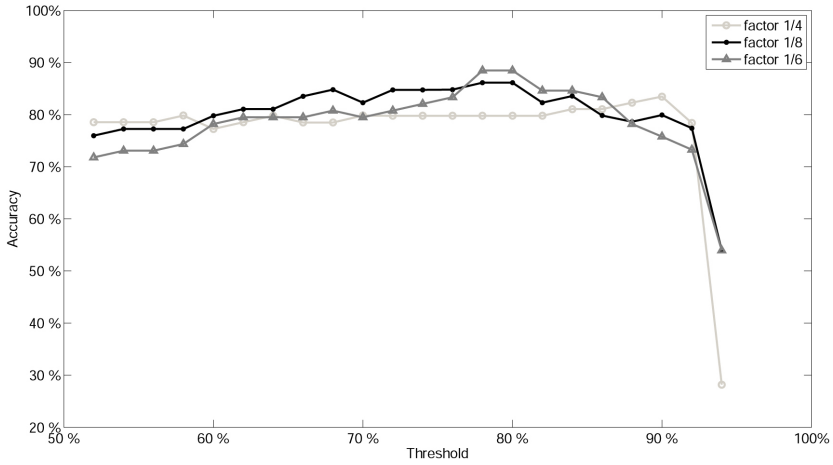

Figura 8.4: Precisión para del sistema basado en SVM con umbral T variable y para distintos factores de reducción.

CAPÍTULO 9

Componentes Principales

El Análisis de Componentes Principales (en inglés, Principal Component Analysis (PCA)) es una técnica estándar para extraer las características más significativas de un conjunto de datos. Se basa en la acción de una transformación lineal (también conocida como la tranformación de *Karhunen-Loéve*) sobre un conjunto de datos de media nula, que diagonaliza su matriz de covarianza. Matemáticamente se define como una transformación lineal ortogonal que transforma los datos en un nuevo conjunto de variables que agrupan la mayor cantidad de varianza, denominadas Componentes Principales (CP). La primera componente principal contendrá las características de los datos con mayor contribución a la varianza, seguida por orden decreciente en su valor de la varianza por la segunda componente principal, tercera, etc... Existen varias construcciones equivalentes entre ellas, que conducen a la obtención de estas Componentes Principales, siendo cada una más apropiada en un contexto diferente.

9.1. Análisis de Componentes Principales

El conjunto de datos formado por n imágenes cerebrales 3D $\mathbf{\Gamma}_i$, cuyo tamaño típico es de $m = 79 \times 95 \times 69 \sim 5 \cdot 10^5$ voxels, se entenderá en este contexto como un conjunto de vectores columna $\mathbf{\Gamma}_i \in \mathbb{R}^m$, $i = 1, 2, ..., n$, formados por la concadenación de los voxels de la imagen. Así, $\mathbf{\Gamma}_i^T = (\Gamma_1, \Gamma_2, ..., \Gamma_m)_i$, donde Γ_j representa el valor de la intensidad correspondiente al voxel j.

Tranformación de Karhunen-Loéve: Sea $\mathbf{\Gamma} \in \mathbb{R}^m$ un vector m-dimensional, existe una representación exacta de éste a través de un conjunto de m vectores linealmente independientes $\mathbf{e}_i \in \mathbb{R}^m$ como:

$$\mathbf{\Gamma} = \sum_{i=1}^{m} z_i \mathbf{e}_i \tag{9.1}$$

donde se asume que los vectores $\mathbf{e}_j$ están sujetos a la condición de ortogonalidad:

$$\mathbf{e}_i^T \mathbf{e}_j \;=\; \delta_{ij} \tag{9.2}$$

donde δ_{ij} es la delta de Kronecker. De esta manera, la ecuación (9.1) no describe otra cosa que un cambio de coordenadas a una nueva base ortonormal de $\mathbb{R}^m$, donde las coordenadas del vector $\mathbf{\Gamma}$ en la nueva base vendrán dadas por:

$$z_i = \mathbf{e}_i^T \mathbf{\Gamma} \tag{9.3}$$

A esta coordenada z_i la llamaremos la componente i-ésima en el nuevo espacio engendrado por la base $\{\mathbf{e}_i\}$.

Supongamos que, en lugar de una representación fiel de $\mathbf{\Gamma}$ como en eq. (9.1), estamos interesados en aproximar $\mathbf{\Gamma}$ usando un número reducido ($p < m$) de vectores de la base $\{\mathbf{e}_i\}$. Una forma de hacerlo sería sustituir algunas componentes z_i, cuyos valores no calculamos, por constantes arbitrarias b_i, de manera que se construye la siguiente aproximación de $\mathbf{\Gamma}$:

$$\hat{\mathbf{\Gamma}} = \sum_{i=1}^{p} z_i \mathbf{e}_i + \sum_{i=p+1}^{m} b_i \mathbf{e}_i \tag{9.4}$$

El error que se comente al aproximar $\mathbf{\Gamma}$ por $\hat{\mathbf{\Gamma}}$ vendrá dado por:

$$\begin{aligned} \Delta\mathbf{\Gamma} =& \mathbf{\Gamma} - \hat{\mathbf{\Gamma}} \\ =& \sum_{i=1}^{m} z_i \mathbf{e}_i - \sum_{i=1}^{p} z_i \mathbf{e}_i - \sum_{i=p+1}^{m} b_i \mathbf{e}_i = \\ =& \sum_{i=p+1}^{m} (z_i - b_i)\mathbf{e}_i \end{aligned} \tag{9.5}$$

Seguiremos un criterio de mínimos cuadrados para obtener una solución optima al problema de la aproximación, buscado aquel valor de las constantes b_i que minimice el error del cuadrado de la media (Minimum mean-square error (MMSE)):

$$MMSE = E\left\{\Delta\mathbf{\Gamma}^2\right\} = \sum_{i=p+1}^{m} E\left\{(z_i - b_i)^2\right\} \tag{9.6}$$

Por lo tanto minimizar el error en el cuadrado de la media equivale a buscar una solución a:

$$\frac{\partial}{\partial b_i} E\left\{(z_i - b_i)^2\right\} = -2(E\{z_i\} - b_i) = 0 \tag{9.7}$$

que sencillamente conduce a:

$$b_i = E\{z_i\} = \mathbf{e}_i^T E\{\mathbf{\Gamma}\} \tag{9.8}$$

quedando determinadas las constantes b_i a el valor esperado de las componentes z_i. Ahora, se puede reescribir el error en el cuadrado de la media como:

$$
\begin{aligned}
MMSE &= \sum_{i=p+1}^{m} E\left\{(z_i - E\{z_i\})^2\right\} = \\
&= \sum_{i=p+1}^{m} \mathbf{e}_i^T E\left\{(\mathbf{\Gamma} - E\{\mathbf{\Gamma}\})\right\} E\left\{(\mathbf{\Gamma} - E\{\mathbf{\Gamma}\})\right\}^T \mathbf{e}_i = \\
&= \sum_{j=p+1}^{m} \mathbf{e}_i^T \Sigma_{\mathbf{\Gamma}} \mathbf{e}_i
\end{aligned} \tag{9.9}
$$

donde $\Sigma_{\mathbf{\Gamma}}$ es, por definición, la matriz de covarianza de $\mathbf{\Gamma}$. Se puede demostrar (ver (Fukunaga, 1990; Miranda et˜al., 2008)) que la elección optima para $\mathbf{e}_i$ es aquella que satisface:

$$
\Sigma_{\mathbf{\Gamma}} \mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{9.10}
$$

o dicho de otro modo, aquella en la que $\mathbf{e}_i$ y λ_i son los autovectores y autovalores de la matriz de covarianza. La expansión de un vector en términos de autovectores de la matriz de covarianza se denomina expansión de Karhunen-Loéve.

9.2. Reducción de la dimensionalidad mediante selección de CPs

Un caso real de base de datos de imagenes cerebrales contendrá un número de imagenes del orden de magnitud de la centena ($n \sim 100$ (Ishii et˜al., 2006)). El valor esperado y la matriz de covarianza habrán de ser estimados por la media muestral:

$$
\mu = \frac{1}{n} \sum_{i=1}^{n} \mathbf{\Gamma}_i \tag{9.11}
$$

y la covarianza muestral:

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{\Gamma}_i - \mu)(\mathbf{\Gamma}_i - \mu)^T \tag{9.12}$$

De la ecuación (9.8), se deduce que la transformación de Karhunen-Loéve de los datos originales se simplifica si se centran los datos extrayendose la media muestral, de manera que se elimina el segundo término en (9.4) que no contiene información relevante sobre la varianza de los datos. Se ha demostrado que, más que una simplificación, trabajar con datos de media nula es una condición necesaria para la obtención de las componentes principales (Miranda et˜al., 2008). Por lo tanto, la transformación de PCA actuará sobre el nuevo conjunto:

$$\mathbf{\Phi}_i = \mathbf{\Gamma}_i - \mu \;\; i = 1, 2, ..., n \tag{9.13}$$

y estará compuesta por un conjunto de m autovectores ortogonales $\mathbf{e}_i$ de la matriz de covarianza muestral:

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{\Phi}_i\mathbf{\Phi}_i^T = \frac{1}{n-1}\mathbf{A}\mathbf{A}^T \tag{9.14}$$

donde $\mathbf{A} = [\mathbf{\Phi}_1, \mathbf{\Phi}_2, ...\mathbf{\Phi}_n]$, tales que:

$$\mathbf{z}_i = \mathbf{e}_i^T\mathbf{A} \tag{9.15}$$

Denominaremos a este vector fila $\mathbf{z}_i \in \mathbb{R}^n$ la i-ésima componente principal. Existirán m componentes principales, entre las que no exisitirá correlación ya que su matriz de covarianza será diagonal. Los autovalores de la matriz de covarianza λ_i nos darán la varianza de las componentes principales ya que:

$$\mathbf{E}^T\Sigma_{\mathbf{\Phi}}\mathbf{E} = \Sigma_{\mathbf{Z}} = \text{diag}(\lambda_1^2, \lambda_2^2, ..., \lambda_m^2) \tag{9.16}$$

donde $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_m]$ y $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m]$. De las propiedades de la traza, se deduce tambien que:

$$\mathrm{tr}(\Sigma_{\mathbf{\Phi}}) = \mathrm{tr}(\Sigma_{\mathbf{Z}}) = \sum_{i=1}^{n} \lambda_i^2 \tag{9.17}$$

lo que permite dar una interpretación a la ecuación de minimización del error en el cuadrado de la media (9.9):

$$MMSE = \sum_{i=p+1}^{m} \mathbf{e}_i^T \Sigma_{\mathbf{\Gamma}} \mathbf{e}_i = \sum_{i=p+1}^{m} \Sigma_{\mathbf{Z}} = \sum_{i=p+1}^{m} \lambda_i^2 \tag{9.18}$$

El proceso de optimización buscará por tanto, los $m - p$ valores de λ_i cuya suma sea mínima, o dicho de otra manera, consistirá en seleccionar aquellas p componentes principales cuya varianza sea máxima. Así, la aproximación (9.4) vendrá dada por la combinación lineal de los p autovectores $\mathbf{e}_i$ cuyas componentes principales tengan mayor varianza, recogiendose las características de mayor variabilidad de los datos en un número $p < m$ de variables:

$$\hat{\mathbf{A}} = \sum_{i=1}^{p} \mathbf{e}_i \mathbf{z}_i \tag{9.19}$$

En resumen, una transformación de PCA consistirá en la diagonalización de la matriz de covarianza de los datos centrados mediante un conjunto de autovectores ortonormales, lo cual es siempre posible debido al teorema de descomposición espectral, ya que es una matriz simétrica definida positiva. Una vez diagonalizada, se seleccionarán los autovalores más altos y sus correspondientes autovectores, que se usarán para representar las características de los datos reduciendo los grados de libertad del sistema.

9.2.1. Eigenbrains

En correspondencia con la terminología usada en el campo del reconocimiento de caras, dónde se usa el término *eigenfaces* para denominar a los autovalores de la matriz de covarianza (Turk and Pentland, 1991), llamaremos '*eigenbrains*' a los autovectores $\mathbf{e}_i$, por su apariencia de imagen cerebral,

refeririéndonos también al espacio que engendran como '*espacio de eigenbrains*'. Para obtenerlos, será necesario diagonalizar una matriz $m \times m$, que en el caso de imagenes cerebrales se convertirá en una matriz $5 \cdot 10^5 \times 5 \cdot 10^5$. La complejidad computacional del proceso de diagonalización se puede ver reducida si se ataca el problema de diagonalizar la matriz $\hat{\mathbf{C}} = \mathbf{A}^T\mathbf{A}$, cuyo tamaño es $n \times n$, normalmente con $n << m$. Si llamamos $\mathbf{v}_j \in \mathbb{R}^n$ a los autovectores de $\hat{C}$:

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_j = \mu_j\mathbf{v}_j, \qquad i = j, 2, ..., n \tag{9.20}$$

Multiplicando esta ecuación por la izquierda por $\mathbf{A}$, conduce a:

$$\mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{v}_j = \mu_j\mathbf{A}\mathbf{v}_j \tag{9.21}$$

de manera que $\mathbf{A}\mathbf{v}_j$ son autovectores de $\mathbf{A}\mathbf{A}^T$. Esto permite obtener n de los m autovectores $\mathbf{e}_i$ de $\mathbf{C}$ de los autovectores $\mathbf{v}_j$ de $\hat{\mathbf{C}}$ como:

$$\mathbf{e}_j = \mathbf{A}\mathbf{v}_j, \quad j = 1, .., n \tag{9.22}$$

Normalmente, solo un numero reducido de eigenbrains es necesario dar cuenta de la mayor parte de la varianza muestral, por lo que solo un pequeño número p será necesario para describir apropiadamente el conjunto de datos (en este caso $p = n$). A menudo, incluso un subconjunto de estos eigenbrains será suficiente para representar correctamente el conjunto de datos. Sin embargo, no ha quedado demostrado que los n eigenbrains obtenidos a través de $\hat{\mathbf{C}}$ sean suficientes o los adecuados para representar bien la matriz de datos. Existe un argumento que demuestra que en efecto lo son, basado en la relación de PCA con una descomposición de valor singular (Singular Value Decomposition (SVD)). La transformación (9.15) es equivalente a la SVD de la matriz de datos $\mathbf{A}$, que viene dada por:

$$\mathbf{A} = \mathbf{E}\mathbf{D}\mathbf{V}^T \tag{9.23}$$

donde $\mathbf{E}$ y $\mathbf{V}$ son matrices ortogonales $m \times m$ y $n \times n$ respectivamente, y $\mathbf{D}$ es una matriz diagonal $m \times n$. La SVD garantiza que la matriz $\mathbf{D}$

queda únicamente determinada por el valor de $\mathbf{A}$ si los valores de $\mathbf{D}$ estan organizados en orden decreciente, mientras que las matrices $\mathbf{E}$ y $\mathbf{V}$ no quedan completamente determinadas. Podemos expresar la matriz $\mathbf{C}$ en términos de la descomponsición singular como:

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T = \mathbf{E}\mathbf{D}\mathbf{D}^T\mathbf{E}^T \tag{9.24}$$

que no es otra cosa que la ecuación (9.16) con:

$$\mathbf{D}\mathbf{D}^T = \Sigma_{\mathbf{Z}} = \text{diag}(\lambda_1^2, \lambda_2^2, ..., \lambda_m^2) \tag{9.25}$$

Haciendo lo mismo con $\hat{\mathbf{C}}$ se llega a:

$$\hat{\mathbf{C}} = \mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T \tag{9.26}$$

donde ahora:

$$\mathbf{D}^T\mathbf{D} = \text{diag}(\lambda_1^2, \lambda_2^2, ..., \lambda_n^2) \tag{9.27}$$

Puesto que $\mathbf{D}$ es diagonal, la única forma que puede tomar es:

$$\mathbf{D} = \begin{pmatrix} \underbrace{\begin{matrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \lambda_n \end{matrix}}_{n} & \underbrace{\begin{matrix} \dots & 0 \\ \dots & 0 \\ & \vdots \\ \dots & 0 \end{matrix}}_{m-n} \end{pmatrix} \tag{9.28}$$

lo que obliga a que al menos $\lambda_k = 0$ con $k = n+1, ..., m$, garantizando que sólo es necesario calcular[1] los n eigenbrains obtenidos a través de $\hat{\mathbf{C}}$. Por consiguiente, la representación en términos de eigenbrains del conjunto de datos dependerá del tamaño muestral n.

[1]La implementación de PCA que usaremos, basada en la función 'princomp' de Matlab [] hará uso de esta descomposición singular para la obtención de las componentes principales.

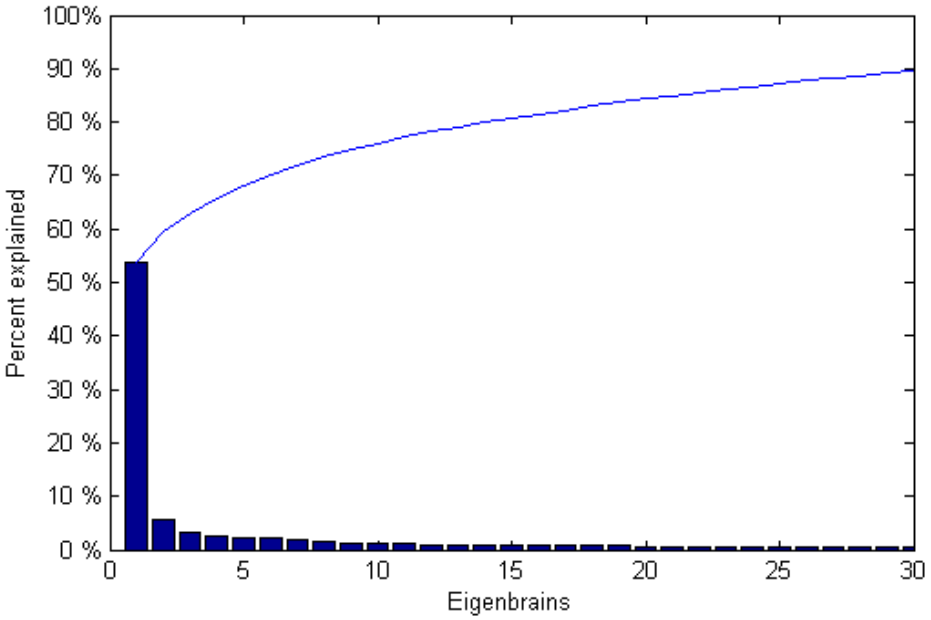

Figura 9.1: Varianza explicada por los 76 primeros eigenbrains

9.2.2. Selección de características a partir de CPs

Como vimos en la sección 4.3.1, la transformación de PCA pertenece a la clase de transformaciones lineales que permiten la reducción dimensional del espacio de características a través de una ecuación del tipo (4.27), una vez establecido que solo $n < m$ eigenbrain entrarán en juego en la ecuación (9.15). A menudo, un número aun más reducido de eigenbrains será necesario para dar cuenta de la mayor parte de la varianza. Como muestra la figura 9.1, en un caso típico de imagenes solo harán falta los 30 primeros eigenbrains de 79 para explicar el 90 % de la varianza.

Sin embargo, la ecuación (9.15) no está en la forma (4.27), ya que las componentes principales son combinaciones lineales de los vectores originales. Se puede reorganizar la información contenida en las componentes principales para que sea útil en el aprendizaje de la siguiente manera: La ecuación (9.19) define la representación del conjunto de vectores imagen en la base de los eigenbrains, cuyas coordenadas vienen dadas por la ecuación (9.15), es decir, las componentes principales. Expandiremos esta última ecuación para

mostrar exactamente como se obtiene el conjunto reducido de p componentes principales:

$$\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_p \end{pmatrix} (\mathbf{\Phi}_1, \mathbf{\Phi}_2, ..., \mathbf{\Phi}_n) \tag{9.29}$$

que, mostrando las coordenadas de cada vector, puede ser expandida a:

$$\begin{pmatrix} (z_1, z_2, ..., z_n)_1 \\ (z_1, z_2, ..., z_n)_2 \\ \vdots \\ (z_1, z_2, ..., z_n)_p \end{pmatrix} = \begin{pmatrix} (e_1, e_2, ..., e_m)_1 \\ (e_1, e_2, ..., e_m)_2 \\ \vdots \\ (e_1, e_2, ..., e_m)_p \end{pmatrix} \left(\begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_m \end{pmatrix}_1, \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_m \end{pmatrix}_2, \cdots, \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_m \end{pmatrix}_n \right) \tag{9.30}$$

Uno puede ver esta ecuación de otra manera, sin más que trasponer a ambos lados de la ecuación, obteniendo un nuevo conjunto de vectores de características:

$$\mathbf{x}_j = \mathbf{\Phi}_j^T \tilde{\mathbf{E}} \quad j = 1, ..., n \tag{9.31}$$

donde $\tilde{\mathbf{E}} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p]$. Este nuevo conjunto de vectores $\mathbf{x}_j \in \mathbb{R}^p$, donde p es el número de eigenbrains seleccionado, estará formado por una reorganización de las componentes principales, pero ahora consiguiendo que cada vector $\mathbf{x}_j$ mantenga su correspondiente etiqueta de clase $y_i \in \{\pm 1\}$, pudiendo ser usados para clasificación. Ahora, la información contenida en las componentes principales se organiza de manera que un vector $\mathbf{x}_i$ se puede interpretar geométricamente como la proyección de una imagen cerebral $\mathbf{\Phi}_i$ en el subespacio de los p eigenbrains más relevantes, dada por sus coordenadas en este espacio. De esta manera, PCA es una herramienta poderosa para conseguir la reducción de la dimensionalidad del espacio de características, pasando de ser $\mathbb{R}^m$ a $\mathbb{R}^p$, donde $m \sim 10^6$, $n \sim 10^2$ y $p \leq n$.

9.2.3. Selección mediante el criterio de Fisher

Seleccionar un subconjunto de eigenbrains para representar las imagenes según su varianza es un método efectivo y simple para reducir la dimensionalidad del espacio de características. Sin embargo, es posible que las características de mayor varianza no sean las mejores para distinguir entre clases, ya que puede haber variabilidad en los datos que responda a factores no relacionados con la enfermedad a diagnosticar y sean comunes en ambas clases.

Para eliminar esta posibilidad, se pueden diseñar criterios de selección de componentes principales que recojan la información que mejor distingue entre clases. Un ejemplo es usar el criterio del factor discriminante de Fisher (Fisher Discriminant Ratio (FDR) (Fisher, 1936)). El FDR se define como:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{9.32}$$

donde μ_i y σ_i^2 son la media y la varianza muestral de la clase $i = 1, 2$. El criterio de selección del factor discriminante de Fisher consiste en usar el FDR como valor para la elección de componentes en lugar de la varianza. Siguiendo este criterio, se extraerá un subconjunto de q componentes principales y sus correspondientes eigenbrains, cuyo FDR toma un valor máximo, permitiendo una construcción diferente de los vectores de características (9.31).

9.3. Filtro de Componentes Principales

Usando el criterio del factor discriminante de Fisher se descubre que una pequeña cantidad de eigenbrains contiene información ruidosa que empobrece la capacidad de distinción entre clases. Esto sugiere el uso combinado de PCA y FDR como filtro para el preprocesamiento de las imagenes. Mediante este filtro se conseguiría eliminar parte de la información contenida en las imágenes originales $\mathbf{\Phi}_j$ que es irrelevante para la clasificación, pudiéndose usar posteriormente otros métodos de extracción de características para definir los vectores de entrenamiento y test a partir de estas imágenes filtradas. Explícitamente, las imagenes filtradas $\hat{\mathbf{\Phi}}_j$ se construirían sustrayen-

do de las imagenes originales la información contenida en un subconjunto q de componentes principales, cuyo FDR es *mínimo*:

$$\hat{\mathbf{A}} = \mathbf{A} - \sum_{i=1}^{q} \mathbf{e}_i \mathbf{z}_i \tag{9.33}$$

donde $\hat{\mathbf{A}} = [\hat{\mathbf{\Phi}}_1, \hat{\mathbf{\Phi}}_2, ..., \hat{\mathbf{\Phi}}_n]$. También se puede entender la tranformación (9.33) como una reconstrucción de las imágenes originales $\mathbf{A}$ usando los $n-q$ eigenbrains cuyo FDR es *máximo*:

$$\hat{\mathbf{A}} = \mathbf{A} - \sum_{i=1}^{q} \mathbf{e}_i \mathbf{z}_i = \sum_{i=1}^{n} \mathbf{e}_i \mathbf{z}_i - \sum_{i=1}^{q} \mathbf{e}_i \mathbf{z}_i = \sum_{i=1}^{n-q} \mathbf{e}_i \mathbf{z}_i \tag{9.34}$$

Es importante observar que, usada de esta manera, la transformación de PCA no supone una reducción en la dimensionalidad del espacio de características, ya que $\hat{\mathbf{A}}$ y $\mathbf{A}$ tendrán la misma dimensión.

9.4. Eliminación de Correlación mediante CPs

La transformación de PCA no solo tiene utilidad como método de reducción de la dimensionalidad del espacio de características. Hemos descrito una aplicación de PCA junto con FDR como método de filtrado de imágenes y a continuación describiremos un marco en el que la transformación de PCA se usa como método para *de-correlacionar* los datos. Hemos visto cómo la representación en términos de eigenbrains está limitada por el tamaño muestral n. Esta limitación puede resultar artificial, y puede conducir a resultados más o menos sesgados en función de la naturaleza del conjunto de datos. Para mejorar esta deficiencia propondremos el uso de vectores representativos del conjunto de datos para describir las características de éste. Concretamente, estos vectores representativos $\mathbf{\Psi}_k \in \mathbb{R}^m$ hacen referencia al valor esperado de los vectores pertenecientes a una cierta clasee $\mathcal{C}^k$. Puesto que cada imagen $\mathbf{\Gamma}_i$ lleva asociada una etiqueta $y_i \in \mathcal{C}^k$, se definen estos vectores característicos como:

$$\mathbf{\Psi}_k = [E\{\mathbf{\Gamma}_i\} : y_i \in \mathcal{C}^k], \quad k = 1, 2, ..., K \tag{9.35}$$

donde K es el número total de clases. Los vectores que pertenezcan a una misma clase k compartirán cierta característica y esta característica vendrá representada por $\mathbf{\Psi}_k$. La motivación para el conjunto de $\mathbf{\Psi}_k$ es que éstos definen un subespacio en $\mathbb{R}^m$, que denominaremos *subespacio representativo*, donde las propiedades de las diferentes clases quedan reflejadas. Por lo tanto, trabajando en este subespacio se consigue una reducción de la dimensionalidad sin pérdida de la información relevante para clasificación.

En el caso del diagnóstico del Alzheimer, las diferentes clases vendrán determinadas por las etapas de la enfermedad. En el caso más sencillo se definirán 2 clases $\mathcal{C}^1$ y $\mathcal{C}^2$, correspondientes a *Normal (NOR)* o *demencia tipo Alzheimer (DTA)*, siendo:

$$\begin{array}{l} y_i = -1 \longrightarrow y_i \in \mathcal{C}^1 \\ y_i = +1 \longrightarrow y_i \in \mathcal{C}^2 \end{array} \tag{9.36}$$

Sin embargo, el Alzheimer es una enfermedad degenerativa, que paulatinamente va afectando diferentes regiones cerebrales, por lo que podrán definirse clases intermedias que impliquen un deterioro cognitivo leve (en ingés, Mild congitive impairment (MCI)) previo a la demencia. Los vectores $\mathbf{\Psi}_k$ serán imagenes representativas de las diferentes etapas, donde cada uno reflejará aquellas regiones del cerebro que se ven afectadas en el proceso degenerativo.

En la práctica, los vectores representativos $\mathbf{\Psi}_k$ deberán ser estimados por las medias muestrales intra-clase $\mathbf{r}_k \in \mathbb{R}^m$, que se definen como:

$$\mathbf{r}_k = \frac{1}{N_k} \sum_{\{y_i \in \mathcal{C}^k\}} \mathbf{\Gamma}_i \ , \quad k = 1, 2, ..., K \tag{9.37}$$

donde N_k denota el número de imágenes en la clase $\mathcal{C}^k$. Para usar la información contenida en estos vectores para clasificación, usaremos una base de $\mathbb{R}^m$ que contenga a estos K vectores para describir el conjunto de vectores imagenes, como los K primeros elementos de la base. Solo estaremos interesados

en las coordenadas de cada vector imágen correspondientes estos vectores, por lo que la base podrá ser completada con $m - K$ vectores arbitrarios. La base a utilizar será por tanto:

$$B = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_K, \mathbf{u}_{K+1}, ..., \mathbf{u}_m\} \tag{9.38}$$

donde $u_i,\ i = K + 1, ..., m$ son vectores arbitrarios (por ejemplo una base cartesiana como en 4.28)[2]. En esta base, las K primeras coordenadas $\mathbf{t}_i = (t_1, t_2, ..., t_K)_i$ de cualquier vector imágen i vendrán dadas por:

$$\mathbf{t}_i = \mathbf{R}^T \mathbf{\Phi}_i,\ i = 1, 2, ..., n \tag{9.39}$$

donde $\mathbf{R}^T = \mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_K$. Esta información puede usarse como vector de características para clasificación. Sin embargo, si las diferencias entre clases son sutiles, los vectores representativos $\mathbf{r}_k$ estarán lejos de ser linealmente independientes. Esto hará que los K primeros vectores de la base B formen ángulos muy pequeños entre ellos, en el sentido Euclídeo, produciendo que las K primeras coordenadas de cualquier vector en esta base sean práctimente iguales, es decir, esten fuertemente correlacionadas (ver fig. 9.5).

Este hecho justifica aquí el uso de PCA, no como técnica de reducción dimensional, sino como método para obtener un nuevo conjunto de K vectores de-correlacionados. Según la ecuación (9.15), la transformación de PCA aplicada a $\mathbf{R}$, definido previamente como el conjunto de vectores representativos, conducirá a un nuevo conjunto de K variables o componentes principales, cuya matriz de covarianza es diagonal, es decir, estan decorrelacionadas:

$$\mathbf{z}_i = \mathbf{e}_i^T \mathbf{R} \tag{9.40}$$

donde $\mathbf{z}_i$ son las $i = 1, ..., m$ componentes principales y $\mathbf{e}_i$ los $i = 1, ..., m$ autovectores de la matriz de covarianza $\Sigma_{\mathbf{R}}$. Como quedó demostrado anteriormente, sólo K de estos m autovectores podrán tener autovalor distinto de

[2]En caso de que esta complección de la base tuviera vectores que no fueran linalmente independientes de los K primeros, siempre se podría construir otra base formada por vectores linalmente independientes, a través del proceso de Gram-Schmidt

cero, por lo tanto solo existirán K componentes principales. Al igual que hicimos con los vectores representativos en (9.38), podremos usar los eigenbrains obtenidos como los K primeros vectores de una nueva base C de $\mathbb{R}^m$:

$$C = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K, \mathbf{u}_{K+1}, ..., \mathbf{u}_m\} \tag{9.41}$$

para describir el resto de la base de datos. Sin embargo, tendremos la ventaja de que estos nuevos vectores $\mathbf{e}_i$ serán ortogonales, junto con el hecho de que las primeras coordenadas de los vectores característicos en esta base están decorrelacionadas. Esto implica que cualquier vector imagen que guarde semejanza con alguno de los vectores representativos, tendrá unas coordenadas en esta base que estarán, en principio, decorrelacionadas de las coordenadas de un vector semejante a *otro* vector representativo de una clase diferente. Con esto se conseguirá, potencialmente, una separación mejor entre clases. Las primeras K coordenadas $\mathbf{x}_i = (x_1, x_2, ..., x_K)_i$ de un vector de la base de datos C se usarán para clasificación, y vendrán dadas por:

$$\mathbf{x}_i = \mathbf{\Gamma}_i^T \mathbf{E} \quad i = 1, ..., n \tag{9.42}$$

con $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K]$. La principal ventaja de usar PCA sobre los vectores representativos $\mathbf{r}_k$ para construir los vectores de características $\mathbf{x}_j$, es que se reduce el sesgo producido al usar toda la base de datos para construir la transformación de PCA. Las singularidades de imágenes individuales de la muestra concreta se suavizan mediante la introducción de imágenes medias muestrales, a la vez que se potencia la separación de clases a través de las diferencias más significativas. PCA sirve pues, para obtener las características representativas de cada clase, independientes de las demás, sin necesidad de introducir ninguna información específica a cerca de la enfermedad. En este caso, se tienen en cuenta *todos* los eigenbrains, y no sólo los de mayor varianza, y la reducción de la alta dimensionalidad del espacio de características se consigue a través del método de proyección definido a través de una ecuación del tipo (4.26).

Nótese que, si el vector $\mathbf{A}$ sigue una distribución Gaussiana, entonces puede ser transformado via PCA (9.15) a un vector independiente $\mathbf{Z}$ normalmente distribuido y decorrelacionado, cuya matriz de covarianza vendrá dada

por (9.25). Puesto que la matriz $\Sigma_{\mathbf{Z}}$ será diagonal, la densidad de probabilidad conjunta $p(\mathbf{Z})$ se factoriza:

$$\begin{aligned} p(\mathbf{Z}) =& \frac{1}{(2\pi \prod_{i=1}^{n} \lambda_i)^{n/2}} \exp\{-\frac{1}{2}(\mathbf{Z}-\mu_{\mathbf{Z}})^T \Sigma_{\mathbf{Z}}(\mathbf{Z}-\mu_{\mathbf{Z}}))\} = \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\lambda_i)^{n/2}} \exp\{-\frac{1}{2}\frac{\mathbf{z}_i-\mu_i}{\lambda_i}\} = \\ &= \prod_{i=1}^{n} p_{N(\mu_i,\lambda_i)}(\mathbf{z}_i) \end{aligned} \tag{9.43}$$

Por lo tanto, en el caso de un vector con distribución de probabilidad Gaussiana, la proyección en el espacio de eignebrains, acompañada de una rotación y reescaleo del sistema de coordenadas, no solo decorrelaciona el vector, sino que consigue también la *independencia estadística*.

9.5. Experimentos

Realizaremos 2 experimentos diferentes con PCA, cada uno diseñado para utilizar los métodos anteriormente descritos: el primero hará uso de PCA para reducir la dimensionalidad del espacio de características, mientras el segundo hará uso de PCA para decorrelacionar un subconjunto de vectores. Usaremos la base de datos de SPECT del hospital Virgen de las Nieves (ver sección 6.2) para el primer de los experimentos, mientras que usarmos la base de datos de ADNI para el segundo de ellos (ver sección 6.1).

9.5.1. PCA

En este experimento, el tamaño de las 79 imágenes SPECT fue reducido por un factor $1/n^3$, con n tomando valores de 2 a 12. De esta manera conseguimos reducir el efecto de posibles adquisiciones defectuosas en algunas regiones cerebrales, mientras que simplificamos el cómputo de los eigenbrains. Se obtuvo el conjunto de eigenbrains a partir de (9.10, donde solo 76 eigenbrains fueron necesarios para dar cuenta del 99.99 % de la varianza muestral.

Una vez obtenidos se hizo uso de la ecuación (9.31) para obtener la reducción de la dimensionalidad de los vectores de características, haciendo variar p, es decir el número de eigenbrains usados, de $p = 1, 2, ...$ hasta 76. Se siguieron dos criterios determinar el orden en que los eigenbrains serían seleccionados:

- Se seleccionaron los eigenbrains en orden decreciente de varianza, siendo el primero aquel con mayor varianza.
- Se seleccionaron los eigenbrains en orden decreciente de su Fisher Discriminant Ratio (FDR), siendo el primero aquel cuyo FDR entre los valores del vector de características de todos los pacientes fuese mayor entre los pacientes de entrenamiento.

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{9.44}$$

donde μ_i y σ_i^2 denotan la media y la varianza de los vectores de características uni-dimensionales de la clase i respectivamente, formados a partir de cada eigenbrain concreto.

Una vez extraidas las características para la clasificiación, se entrenó con estos datos un SVM usando 4 kernels diferentes: lineal, cuadrático, RBF y polinomial. El test se realizó siguiendo la técnica de validación cruzada de dejar uno fuera, que iterativamente separa un sujeto para el test, mientras el clasificador es entrenado con los restantes datos, de manera que cada sujeto queda fuera una vez (ver 4.5.3).

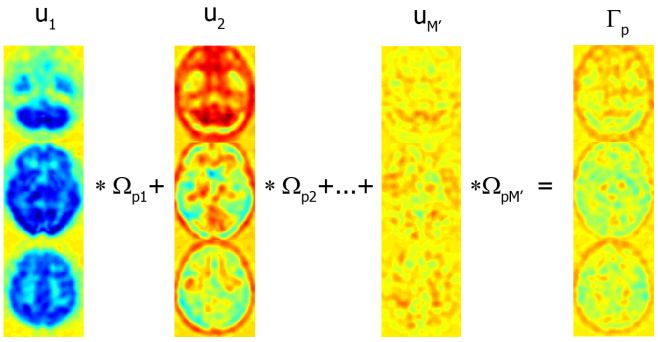

Figura 9.2: Representación en términos de Eigenbrains de la imagen p con media cero. Se muestran tres cortes transaxiales respresentativos.

9.5.2. PCA como proyección

En este experimento se sigue el método de PCA para obtener decorrelación entre imágenes representativas. Para obtener las imágenes representativas muestrales a partir de (9.37), se dividirán los datos en tres grupos diferentes. Cada grupo dará lugar a la construcción de un número diferente de imágenes representativas:

- **Group 1** Se consideran todas las imágenes de la base de datos. Se forman 3 imágenes representativas, una correspondiente a NC, otra a MCI y otra a AD.
- **Group 2** Se considera un subconjunto de los datos. No se tienen en cuenta los sujetos MCI, por lo que se forman 2 imágenes representativas AD y NC.
- **Group 3** e considera un subconjunto de los datos. No se tienen en cuenta los sujetos AD, por lo que se forman 2 imágenes representativas MCI y NC.

Una vez obtenidas las imágenes representativas en cada caso, se aplica la tranformación de PCA (9.40) para extraer los vectores de características, haciendo uso de la base (9.41). Se realizan también pruebas con los vectores (9.39), con el fín de probar la necesidad de aplicar un método de decorrelación sobre la base (9.38). Los vectores de características son usados para entrenar un clasificador SVM usando 2 kernels diferentes: uno lineal, y otro no-lineal (RBF). El test se embebe dentro de un proceso de validación cruzada basado en la técnica de dejar uno fuera, con el fín de estimar los parametros de eficiencia de la clasificación.

9.6. Resultados

9.6.1. PCA

Es de esperar que los kernel lineales tengan un mejor rendimiento que otros cuando la dimensión del espacio de características aumenta. Si este

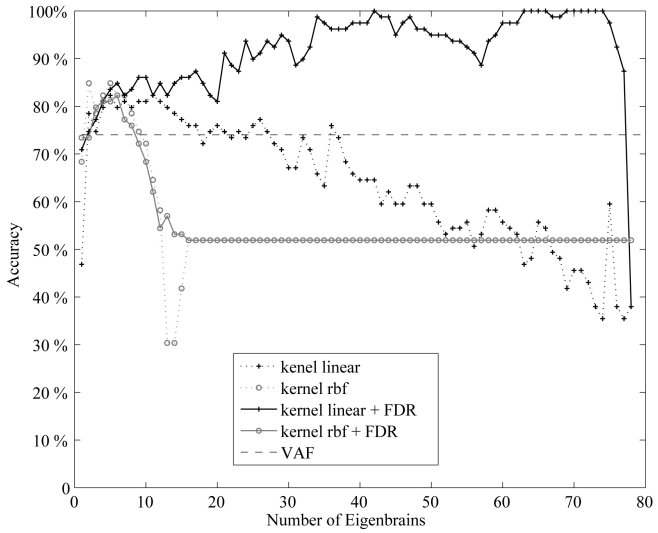

Figura 9.3: Precisión de dos clasificadores usando diferentes kernels frente al numero de eigenbrains usado para contruir los vectores de características para los metodos de seleccionar los eigenbrains.

aumento se realiza siguiendo el criterio de ordenación según la varianza, la información relevante está contenida en los primeros eigenbrains, y el resto sólo contribuye como información ruidosa (véase fig.. 9.3). Pero si el aumento de la dimensión del espacio de características se lleva a cabo siguiendo criterio de ordenación según el FDR, la eficacia en la clasificación aumenta, si se compara con el anterior criterio de ordenación, lo que permite concluir que sólo una pequeña fracción de eigenbrains contiene información ruidosa. Notablemente, se llega a una precisión del 100 % al combinar SVM con kernel lineal y un gran numero de eigenbrains, aumentando la dimensión del espacio de características, independientemente del factor de reducción. Esto supera a los resultados anteriores, como el 74 % de precisión obtenida mediante VAF (Fung and Stoeckel, 2007; Stoeckel et˜al., 2001).

Con nuestro conjunto de datos, se obtuvieron los mejores resultados usando un factor de reducción $1/8^3$ y tomando los tres primeros eigenbrains para construir los datos de entrenamiento y test. En ese caso la precisión alcanzó el 88,6 % y la sensibilidad tomo su segundo valor mas alto 95,6 %. Con estos parametros, la especificidad fue de 81,2 %.

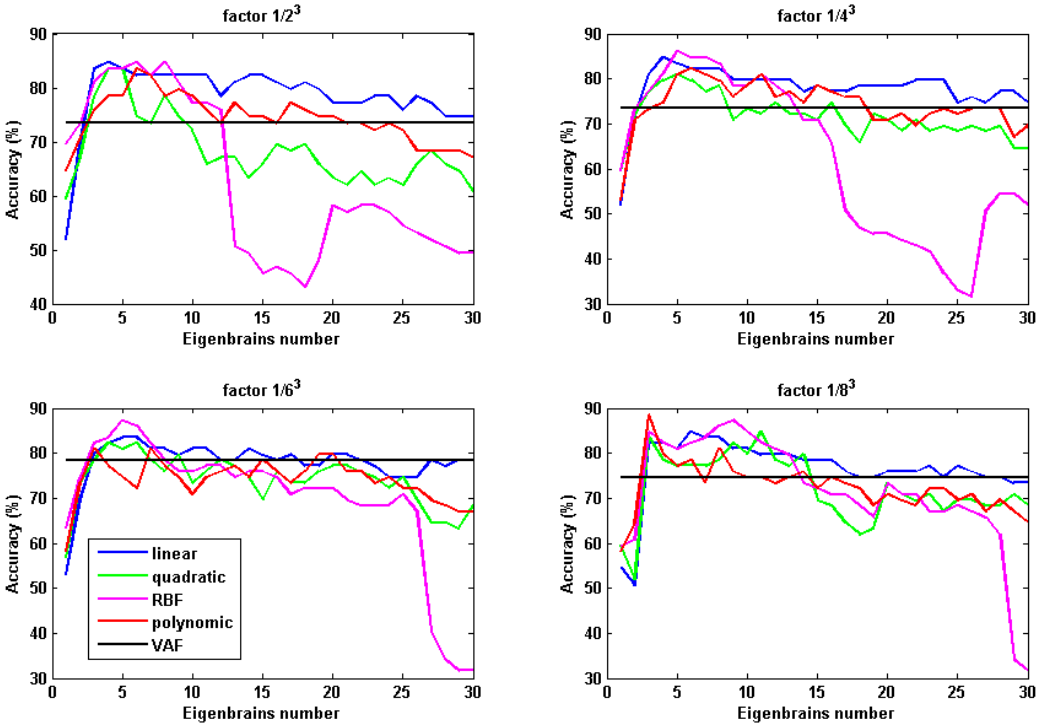

Figura 9.4: Precisión del clasificador para diferentes factores de compresión $f = \{1/2^3, 1/4^3, 1/6^3, 1/8^3\}$ y diferentes kernels frente al número de eigenbrains usado.

9.6.2. PCA como proyección

La tabla 9.1 muestra los resultados de clasificación sobre los tres grupos descritos en la sección 9.5.2. una vez los pasos antes descritos; la proyección, PCA y ICA ha sido aplicada a los datos brutos. Un aspecto importante para la construcción de un clasificador óptimo es la proporción equitativa de muestras de entrenamiento positivas y negativas. El grupo 2 satisface esta condición (con 53 positivos y 52 negativos), pero no el grupo 1 ni el 3 (con 167 positivos y 52 negativos, y 114 positivos y 52 negativos respectivamente) produciendo carencias a la hora de construir el clasificador, y por lo tanto dando como resultado bajos valores de especificidad. Estos valores bajos de especificidad pueden ser tambien causados por la indroducción en los grupos 1 y 3 de pacientes MCI, cuyo patrón de atrofia en las imagenes cerebrales es

	Projection		PCA		
	Linear Kernel	RBF kernel	Linear Kernel	RBF kernel	
Precisión	76.26 %	76.26 %	90.41 %	90.41 %	Group 1
Especificidad	0 %	0 %	75.00 %	71.15 %	
Sensibilidad	100 %	100 %	95.21 %	96.41 %	
Probabilidad positiva	1	1	3.8084	3.3421	
Probabilidad negativa	∞	∞	0.0639	0.0505	
Precisión	60.00 %	53.33 %	84.76 %	84.76 %	Group 2
Especificidad	65.38 %	65.38 %	86.54 %	86.54 %	
Sensibilidad	54.72 %	41.51 %	83.02 %	83.02 %	
Probabilidad positiva	1.5807	1.1992	6.1671	6.1671	
Probabilidad negativa	0.6926	0.8946	0.1962	0.1962	
Precisión	68.67 %	68.67 %	82.53 %	83.13 %	Group 3
Especificidad	0 %	0 %	67.31 %	67.31 %	
Sensibilidad	100 %	100 %	89.47 %	90.35 %	
Probabilidad positiva	1	1	2.7368	2.7637	
Probabilidad negativa	∞	∞	0.1564	0.1434	

Tabla 9.1: Medidas del rendimiento del metodo de PCA para los tres grupos muestrales, en referencia a la simple proyección.

complejo y latamente variable, y evoluciona conforme la enfermedad progresa (Minoshima et˜al., 1997; Drzezga et˜al., 2003; Karas et˜al., 2004).

Con respecto a las diferentes técnicas de extracción de características, la simple proyección funciona mal, cercana a la clasificación al azar como es de esperar. Sin embargo, la aplicación de PCA para decorrelacionar los vectores representativos introduce mejoras significativas, como queda representado en la tabla 9.1. Details are more clearly understood in terms of group 2 results represented in figures 9.5, 9.6 and 10.3, since similar results are found in group 1 and 3, but obscured by the previously mentioned facts.

La figura Fig. 9.5 revela que $\mathbf{\Omega}_1$ y $\mathbf{\Omega}_2$, obtenidos de la ecuación (8.4) como la proyección de cada vector imagen en el subespacio engendrado por los vectores característicos del AD y NC, respectivamente, están altamente correlacionadas, a pesar de que éstas imágenes medias muestran diferencias en la intensidad de algunas regiones del cerebro, como el cingulo posterior gyri. Esto permite afirmar que las imágenes medias no son lo suficientemente representativas y justifica el uso de las técnicas desarrolladas a continuación

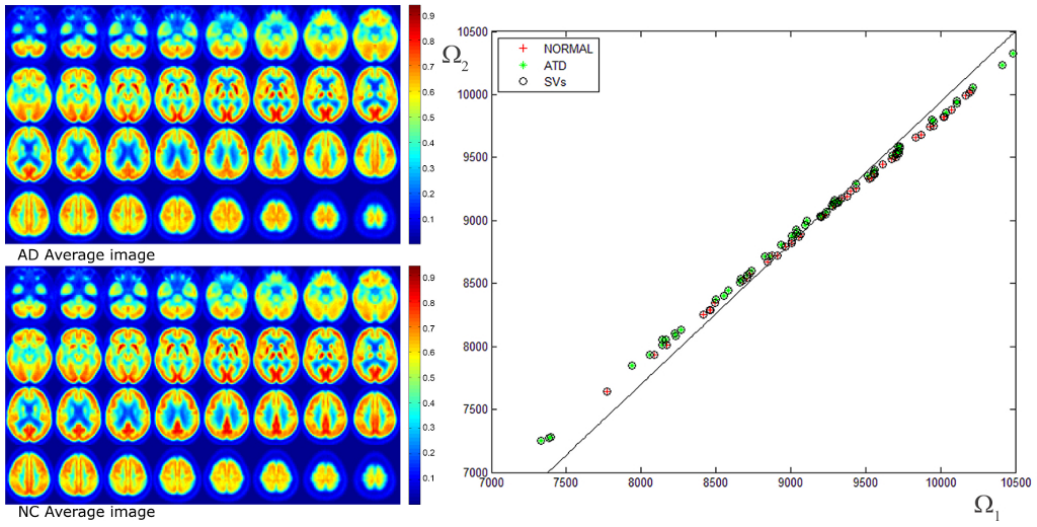

Figura 9.5: Clasificador lineal y vectores de entrenamiento $\mathbf{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio de vectores característicos (izquierda, cortes transaxiales)

para la extracción de una base de vectores que servirá de paleta para describir el resto de las imágenes.

La figura 9.6 muestra cómo la aplicación de PCA a los vectores representativos permite una separación entre clases, y por consiguiente la construcción de un clasificador efectivo. La separación de los vectores de entrenamiento en dos clases diferentes viene dada por el valor que toma $\mathbf{\Omega}_2$, mientras que el valor de $\mathbf{\Omega}_1$ no juega ningún papel en la construcción del clasificador. Concretamente, valores positivos de $\mathbf{\Omega}_2$ corresponden con pacientes afectados por la enfermedad del Alzheimer, mientras que los valores negativos reflejan que el paciente es NC. Si se tiene en cuenta la apariencia de los eigenbrains, se puede entender este hecho de la siguiente manera: el Eigenbrain 1 tiene una apariencia de imagen PET normal, mientras que el Eigenbrain 2 tiene una forma inverosímil, complementaria a la del Eigenbrain 1. El Eigenbrain 1 sirve como base para construir cualquier imagen muestral de la base de datos, contribuyendo con los aspectos básicos de una imagen cerebral funcional. Por otro lado, el Eigenbrain 2 destaca las areas afectadas por la enfermedad que sirven para distinguir entre clases. Específicamente, pone de relieve el cingulo posterior gyri y el precunei, además de la región temporo-parital, ambas con-

sideradas en la literatura como las areas típicas afectadas por hipoperfusión en la enfermedad del Alzheimer (Claus et˜al., 1994; Messa et˜al., 1994; Talbot et˜al., 1998; Minoshima et˜al., 1997). Se observa que también algunas regiones pequeñas del tálamo se seleccionan, lo que no ha sido nunca descrito como regiones relevantes para el diagnóstico. Esto debe interpretarse como un fallo en la normalización espacial: cuando se obtienen los eigenbrains a partir de los vectores relevantes, hay un pequeño desajuste espacial en la zona del talamo. Puesto que tras la aplicación de PCA las nuevas coordenadas agrupan la mayor cantidad de varianza, las regiones adyacentes al tálamo son automáticamente selecionadas para construir el eigenbrain 2. Sin embargo, son lo suficientemente pequeñas como para que su intruducción suponga una cantidad de información despreciable en los vectores de características.

En resumen, una muestra se toma como EA cuando su valor de $\mathbf{\Omega}_2$ es positivo, independientemente de su valor de $\mathbf{\Omega}_1$, complementando de esta manera la imagen del eigenbrain 1 con valores de intensidad menor en el cíngulo, precunei y regiones temporo-parital. Por otro lado, una muestra es considerada como normal cuando su valor de $\mathbf{\Omega}_2$ es negativo, añadiendo a la imagen del eigenbrain 1 un patrón de perfusión normal en las regiones mencionadas anteriormente.

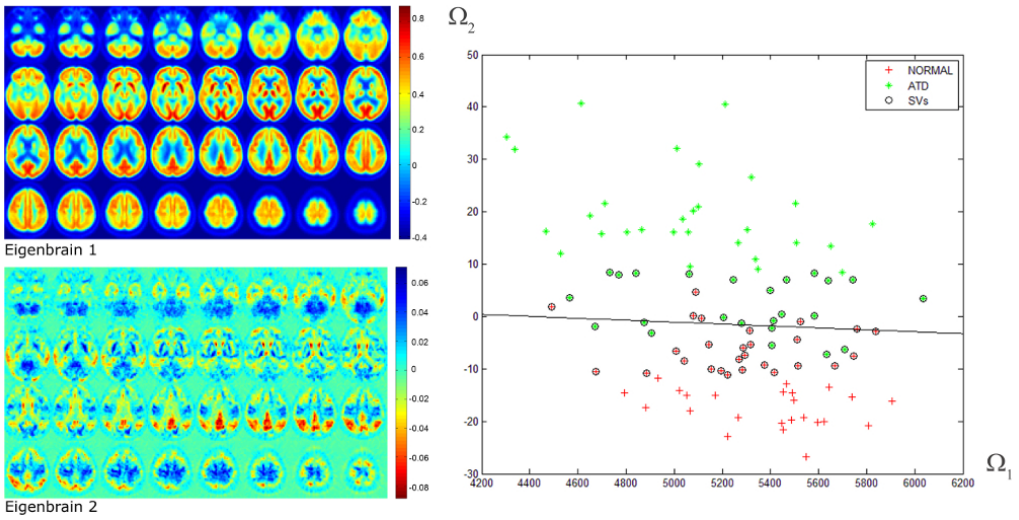

Figura 9.6: Clasificador lineal y vectores de entrenamiento $\mathtt{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio PCA de eigenbrains (izquierda, cortes transaxiales)

CAPÍTULO 10

Componentes Independientes

La independencia estadística es una propiedad deseable en el campo del reconocimiento de patrones, y el Análisis de componentes Independientes (en inglés, Independent Component Analysis (ICA)) proporciona la herramienta necesaria para obtener un conjunto de fuentes independientes a partir de un conjunto de datos. En procesado de señales y procesado de imagenes, el marco habitual en el que se plantea ICA es el problema del 'cocktail party'. En su solución, se recuperan de las fuentes originales que produjeron una señal mezcla registrada, a partir del proceso conocido como separación ciega de fuentes (en inglés Blind Source Separation (BSS)). La resolución de este problema ha de llevarse a cabo sin conocimiento ninguno a cerca de la proporción fuente/ruido que genera la señal registrada. A través de ICA, buscando la independencia estadística de las fuentes originales, se puede obtener el conjunto de fuentes originales, obteniendose varias ventajas sobre otros métodos que también pueden resolver el problema, como PCA. Una de ellas es que las fuentes ruidosas pueden ser separadas completamente y eliminadas, de manera que se obtenga un conjunto independiente de fuentes que contengan toda

la información relevante. Se considera a ICA una generalización de PCA, ya que la independencia estadística se alcanza a todos los ordenes, no solo a primer orden como ocurre con la diagonalización de la varianza en PCA.

10.1. Análisis de Componentes Independientes

La función de ICA consiste en encontrar una solución al problema separación ciega de fuentes (Bell and Sejnowski, 1995; Comon, 1994; Bingham, 2003) sin ruido (dicho de otra manera, no concemos que parte de la fuente es ruido y cual no lo es), que puede ser formulado de la siguiente manera: Sean $\mathbf{A}$ un vector observado y $\mathbf{E}$ una matriz de mezcla de rango completo, el problema consiste en encontrar el conjunto de fuentes $\mathbf{X}$ que, al mezclarse producen el vector observado:

$$\mathbf{A} = \mathbf{ES} \tag{10.1}$$

donde las señales fuente $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$ se asume que son estocásticamente independientes:

$$p_{\mathbf{S}}(\mathbf{u}) = p_{\mathbf{s}_1}(\mathbf{u}_1) \cdots p_{\mathbf{s}_n}(\mathbf{u}_n) \tag{10.2}$$

La finalidad de ICA es encontrar las variables latentes o fuentes $\mathbf{s}_j$ y el proceso de mezcla; en el caso lineal, esto último supone encontrar la matriz de mezcla $\mathbf{E}$. Una manera de hacerlo es buscar la matriz de separación $\mathbf{W}$ de manera que:

$$\mathbf{X} = \mathbf{WA} \tag{10.3}$$

las variables $\mathbf{X}$ son estimaciones de $\mathbf{S}$ sin tener en cuenta reescaleos y permutaciones. Por consiguiente, $\mathbf{W}$ es una estimación de la (pseudo)inversa de $\mathbf{E}$, permitiéndose reescaleos y permutaciones de las filas de $\mathbf{W}$. A menudo, las fuentes se estimarán una a una, a través de un vector columna $\mathbf{w}_j$(que se almacenará como una fila de $\mathbf{W}$) tal que:

$$\mathbf{x}_j = \mathbf{w}_j^T \mathbf{A} \tag{10.4}$$

es una estimación de $\mathbf{s}_j$. Como vimos al final de capítulo 9.4, si $\mathbf{A}$ sigue una distribución gaussiana, existirá una tranformación como (10.3) que lo separe en componentes independientes, en la que las columnas de $\mathbf{W}$ seán sencillamente los autovectores de la matriz de covarianza $\Sigma_{\mathbf{X}}$ de $\mathbf{X}$. Sin embargo, éste no será el caso general, por lo que será necesario diseñar una estrategia para obtener $\mathbf{W}$.

10.1.1. Información mutua

Existen varias propuestas para estimar las componentes independientes, dando lugar a diferentes algoritmos, algunas de las cuales discutiremos brevemente aqui. Empezando desde un punto de vista basado en la teoría de la información, el problema de ICA puede ser formulado en términos de la información mutua entre variables, que será un denominador común a todas las formulaciones del problema. Ésta, viene definida por:

$$I(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}) - H(\mathbf{A}|\mathbf{B}) \tag{10.5}$$

donde $H(\mathbf{A})$ es la entropía de $\mathbf{A}$, mientras que $H(\mathbf{A}|\mathbf{B})$ es la entropía de $\mathbf{A}$ que no proviene de $\mathbf{B}$. Para evitar complejidades[1], será mas conveniente trabajar con gradientes de variables de teoría de la información, por lo que se define la entropía diferencial como:

$$H(\mathbf{x}) = \int p_{\mathbf{x}}(\mathbf{u}) \log p_{\mathbf{x}}(\mathbf{u}) d\mathbf{u} \tag{10.6}$$

donde $p_{\mathbf{x}}$ es la función distribución de probabilidad de $\mathbf{x}$. En el caso de independencia estadística, la distribución de probabilidad conjunta de $\mathbf{X}$ se factorizará, esto es:

[1] Ocurre que el segundo término de (10.5) es mínimo en la ausencia de ruido, divergiendo a $-\infty$.

$$p_{\mathbf{X}}(\mathbf{u}) = \prod_{i=1}^{n} p_{\mathbf{x}_i}(u_i) \tag{10.7}$$

Por lo tanto, una manera natural de comprobar si las estimaciones de las fuentes $\mathbf{X}$ son independientes será medir la distancia entre las funciones de ambos lados de la ecuación (10.7):

$$d\left(p_{\mathbf{X}}(\mathbf{u}), \prod_{i=1}^{n} p_{\mathbf{x}_i}(u_i)\right) \tag{10.8}$$

Usaremos aquí como medida de la 'distancia' $d(\cdot,\cdot)$ la divergencia de Kullback (que no es propiamente una distancia en el sentido tradicional, ya que no se trata de una función simétrica), definida por:

$$d\left(p_{\mathbf{x}}, p_{\mathbf{x}'}\right) = \int p_{\mathbf{x}}(\mathbf{u}) \log \frac{p_{\mathbf{x}}(\mathbf{u})}{p_{\mathbf{x}'}(\mathbf{u})} d\mathbf{u} \tag{10.9}$$

de manera que la divergencia de Kullback satisface:

$$d\left(p_{\mathbf{x}}, p_{\mathbf{x}'}\right) \geq 0 \tag{10.10}$$

saturando la igualdad si y solo si $p_{\mathbf{x}}(\mathbf{u}) = p_{\mathbf{x}'}(\mathbf{u})$. Usaremos esta medida de la distancia precisamente porque, aplicada a (10.7), nos da la información mutua:

$$I(p_{\mathbf{X}}) = \int p_{\mathbf{X}}(\mathbf{u}) \log \frac{p_{\mathbf{X}}(\mathbf{u})}{\prod_i p_{\mathbf{x}_i}(\mathbf{u})} d\mathbf{u} \tag{10.11}$$

que se anulará en el caso de que las $\mathbf{x}_i$ sean mutuamente independientes.

10.1.2. Función de contraste

En todas las propuestas para la estimación de ICA, se elige una función de contraste G. G es una función suave, con valores escalares que mide la eficacia

de la estimación de ICA, en el sentido de que mide cómo de independientes son las estimaciones de las fuentes originales. Según el marco elegido, se tomará una u otra función de contraste.

Nos centraremos en el estudio de la negentropía, como un candidato a función de contraste, debido a las interesantes propiedades que posee. La negentropía mide la no-gaussianidad, es decir, la diferencia de entropía entre una distribución de probabilidad Gaussiana y otra cualquiera. Teniendo en cuenta que una distribución Gaussiana viene dada por al función:

$$f_x(\mathbf{u}) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_x)^{1/2}} \exp\{\frac{-1}{2}\mathbf{u}^T \Sigma_x \mathbf{u}\} \tag{10.12}$$

se define la negentropía como:

$$J(p_x) = H(f_x) - H(p_x) \tag{10.13}$$

donde f_x es una distribución de probabilidad Gaussiana con la misma media y varianza que p_x. De esta manera, la negentropía es siempre positiva, y se hace cero sólo en el caso de que p_x sea tambien Gaussiana. Pero lo que es aún mas interesante es que la negentropía es invariante bajo transformaciones lineales de la variable $\mathbf{x}$.

Podemos reescribir ahora la información mutua en función de la negentropía a partir de 10.6, 10.11 y 10.13 como:

$$I(p_{\mathbf{X}}) = J(p_{\mathbf{X}}) - \sum_{i=1}^{n} J(p_{\mathbf{x}_i}) + \frac{1}{2}\log\frac{\prod \Sigma_{\mathbf{X}ii}}{\det \Sigma_{\mathbf{X}}} \tag{10.14}$$

El último término de la ecuación (10.14) desaparecerá si tomamos la variable $\mathbf{x}$ estandarizada (esto es, con media cero y varianza unidad) y decorrelacionada. Es siempre posible conseguir que una variable este estandarizada, mediante un proceso conocido como estandarización (Comon, 1994), y se puede decorrelacionar a través de una transformación de PCA. Por lo tanto consideraremos que los datos han sido sometidos a este paso previo.

10.1.3. Algoritmos para ICA

En la práctica, encontrar una estimación $\mathbf{X}$ de las fuentes originales $\mathbf{S}$ a través de la matriz de mezcla $\mathbf{W}$, supone maximizar una medida de la independencia estadística de las estimaciones $\mathbf{X}$. Una propuesta para la realización de la transformación de ICA $\mathbf{W}$ es la búsqueda de aquella que haga *mínima la información mutua entre las variables* $\mathbf{x}_i$, ya que esta será cero en caso de independencia. La negentropía proporciona una medida de la información mutua (10.14), y está directamente relacionada con la Gaussianidad de la distribución de las variables. En otras palablas, buscar la minima información mutua entre variables es equivalente a buscar la máxima negentropía, ya que la negentropía es invariante bajo tranformaciones lineales, lo que equivale a buscar aquellas direcciónes de no-Gaussianidad.

El marco de la información mutua permite relacionar el método mencionado con otros métodos para estimar las componentes independientes. Uno de ellos consiste en la maximización de la no-Gaussianidad de las componentes. Éste viene motivado por el teorema de límite central, que establece que la media muestral de una secuencia de variables independientes aleatorias con la misma distribución se aproxima a una distribución normal (o Gaussiana) asintóticamente conforme el número de variables aumenta. Por lo tanto, una suma de variables aleatorias estará mas cerca de ser Gaussiana que tomadas independientemente, lo que equivale a afirmar que si una variable se acerca a la distribución Gaussiana, podrá descomponerse en una combinación de variables más independientes. En términos matemáticos, la estimación:

$$\mathbf{x}_i = \mathbf{w}_i^T \mathbf{A} \tag{10.15}$$

será tanto más independiente cuanto menos Gaussiana sea, y por tanto será mejor estimación de s_i. La no-Gaussianidad se mide a menudo en términos de cumulantes, como kurtosis o skewness, para los que existen métodos robustos (Hyvaerinen, 1999).

Otros métodos propuestos hacen uso de metodos tensoriales. Entre ellos los más conocidos son FOBI (first-order blind identification) (Cardoso, 1990) y JADE (joint approximate diagonalization of eigenmatrices) (Cardoso and Souloumiac, 1993). Los tensores son generalizaciones de los operadores lineales, en particular, los cumulantes son generalizaciones de la matriz de co-

varianza. Minimizar los cumulantes de alto orden aproximadamente produce decorrelación a alto orden, y puede por tanto ser usado para resolver ICA. Sin embargo, las propiedades estadísticas de estos tensores son en general inferiores a las de las herramientas presentadas anteriormente, y su aplicación resulta tediosa en alta dimensión (por Aapo˜Hyvaerinen et˜al., 2001).

Como vimos a través de (10.13), la no-gaussianidad también se mide a través de la negentropía, relacionando el método de la maximización de la no-gaussianidad con el basado en la información mutua. Un método conocido que se apoya en este pricipio de maximización de la información mutua es Infomax (Bell and Sejnowski, 1995). En la práctica, es necesaria alguna forma de calcular la negentropía, siendo frecuente calcularla en términos de cumulantes. Sin embargo, usar cumulantes puede dar como resultado malas aproximaciones de la entropía. Principalmente, existen dos razones para ello. La primera es que las estimaciones de cumulantes de alto orden a traves de conjuntos de muestras finitos son muy sensibles a *outliers*: sus valores dependerán solo de un conjunto de observaciones con valores grandes, que posiblemente correpondan con registros erróneos. Esto implica que los outliers determinarán completamente las estimaciones de los cumulantes, haciéndolos inservibles. En segundo lugar, incluso si las estimaciones muestrales de los cumulantes fuesen perfectas, medirán principalmente las colas de la distribución, y se veran poco afectadas por la estructura cercana al centro de la distribución.

Puesto que el uso de cumulantes para aproximar la negentropía se expone a las inestabilidades antes mencionadas, es necesario alguna aproximación mejor a la entropía. Una de ellas es la basada en una entropía máxima aproximativa, que está en la base del desarrollo del algoritmo de FastICA (Oja, 1997; Hyvaerinen, 1999), que será el que tomaremos para nuestros cálculos. Esta aproximación es más exacta que la dada por cumulantes y tiene mejores propiedades estadísticas para muestras finitas. En cambio, mantiene la simplicidad conceptual y computacional de estos metodos, ya que se aproxima la entropía simplemente por la máxima entropía que es compatible con las observaciones de la variable aleatoria $\mathbf{X}$. Para conseguirlo, se deriva una aproximación de primer orden de la densidad que tiene máxima entropía para un conjunto dado de ligaduras, y se usa para derivar aproximaciones a la entropía de $\mathbf{X}$. En la práctica, estas aproximaciones se traducen a aproximar la negentropía por:

$$J(\mathbf{x}_i) \approx c[E\{G(\mathbf{x}_i)\} - E\{G(\nu)\}]^2 \tag{10.16}$$

donde G es practicamente cualquier función no cuadrática, c es una constante irrelevante, y ν es una variable estandarizada al igual que $\mathbf{x}_i$. Si elegimos la función G como $G(\mathbf{x}_i) = \mathbf{x}_i^4$, entonces se obtiene una generalización de la aproximación basada en cumulantes de (Comon, 1994). Por tanto, diferentes elecciones de G (a la que, por simplicidad, tambien se le suele denominar función de contraste) darán lugar a diferentes formulaciones de la aproximación.

La aproximación (10.16) permite construir un algoritmo muy versátil para la obtención de la tranformación de ICA. Estimando cada fuente una a una $\mathbf{x}_i = \mathbf{w}_i^T\mathbf{A}$, se puede construir una aproximación de la función de contraste como:

$$J_G(\mathbf{w}) = c[E\{G(\mathbf{w}_i^T\mathbf{A})\} - E\{G(\nu)\}]^2 \tag{10.17}$$

con esta aproximación, una minización de la información mutua conllevará una minimización de la función:

$$\sum_i J_G(\mathbf{w}_i) \tag{10.18}$$

sujeta a la condición de decorrelación. Esto implicará buscar una extremo de la función $E\{G(\mathbf{w}_i^T\mathbf{A})\}$, problema que se puede formular en términos de un algoritmo iterativo de punto fijo con la siguiente actualización de $\mathbf{w}$:

$$\mathbf{w}_i \longleftarrow E\{\mathbf{A}g(\mathbf{w}_i^T\mathbf{A})\} - E\{g'(\mathbf{w}_i^T\mathbf{A})\}\mathbf{w} \tag{10.19}$$

donde $\mathbf{w}_i$ son las columnas de la matriz de separación $\mathbf{W}$. En la práctica, los valores esperados han de ser sustituidos por las estimaciones muestrales. La función no lineal g se elige de manera que sea la derivada de la función de contraste G no cuadrática, que mide la no-gaussianidad, negentropía o lo que se estime oportuno. El algoritmo fue construido por primera vez para la función de coste de kurtosis, lo que resulta para g un polinomio de tercer grado, aunque posteriormente se han discutido otras elecciones de la función de contraste más robustas, como por ejemplo funciones no polinómicas

como $\log\cosh(\mathbf{x})$ o $exp(-\mathbf{x}^2)$. Tomaremos este modelo de tercer grado por conveniencia, ya que contrariamente a muchos otros algoritmos, la elección de la función de contraste en FastICA no restringe el tipo de componentes independientes que pueden ser estimadas, sino que es solo importante para optimizar el rendimiento del algoritmo.

Antes de ejecutar el algoritmo (10.19), los datos se transforman de tal modo que tengan media cero y se preprocesan por 'whitening': Un primer vector $\mathbf{w}_i$ de norma unidad es elegido al azar. Después de cada paso de iteración (10.19), $\mathbf{w}_i$ es de nuevo normalizado a norma unidad. Se continúa iterando hasta que la dirección de $\mathbf{w}_i$ no cambia significativamente. En el llamado enfoque de deflación, las componentes independiente $\mathbf{s}_j$ se estiman una por una, y que debe garantizarse que las filas $\mathbf{w}_i$ de la matriz de separación son ortogonales. Esto se consigue después de cada paso de iteración (10.19) sustrayendo de la última $\mathbf{w}_J$ las proyecciones de todas las que se habían estimado anteriormente $\mathbf{w}_p,\ p = 1, ..., J-1$. Después el vector $\mathbf{w}_J$ se normaliza de nuevo, pudiendose probar su convergencia solo tras este paso. La convergencia del algoritmo de deflación ha sido demostrado ser cúbica.

10.2. Reduccion de la dimensionalidad mediante ICs

Al igual que con las componentes principales, las componentes independientes obtenidas del proceso de minimización de la información mutua pueden ser usadas para reducir la dimensionalidad del espacio de características, y extraer de los datos originales unos vectores de características que contengan la información relevante. Reescribiendo la estimación de las fuentes independientes (10.15) como:

$$\begin{pmatrix} (x_1, x_2, ..., x_m)_1 \\ (x_1, x_2, ..., x_m)_2 \\ \vdots \\ (x_1, x_2, ..., x_m)_n \end{pmatrix} = \begin{pmatrix} (w_1, w_2, ..., w_n)_1 \\ (w_1, w_2, ..., w_n)_2 \\ \vdots \\ (w_1, w_2, ..., w_n)_n \end{pmatrix} \begin{pmatrix} (\Phi_1, \Phi_2, ..., \Phi_m)_1 \\ (\Phi_1, \Phi_2, ..., \Phi_m)_2 \\ \vdots \\ (\Phi_1, \Phi_2, ..., \Phi_m)_n \end{pmatrix} \tag{10.20}$$

Podría considerarse el caso en el que existiesen más fuentes originales que observadas, aunque nosotros no cosideraremos esa posibilidad, ya que estamos interesados en la reducción de la dimensión del espacio de características. En la ecuación anterior se da el caso en el que se registran el mismo número de fuentes n que vectores observados. Una vez obtenidas las fuentes originales, usaremos el método desarrollado en la sección 4.3.1 para proyectar cada imagen al subespacio engendrado por estas fuentes independientes. Explícitamente, haremos uso de:

$$\mathbf{z}_i = \mathbf{\Phi}_i^T \mathbf{X} \tag{10.21}$$

que está en la forma de la ecuación 4.26, donde $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$. De esta manera conseguiremos reducir el espacio de características de $\mathbb{R}^m$ a $\mathbb{R}^n$. Es posible seleccionar un número menor de fuentes independientes $\mathbf{x}_i$ para formar los nuevos vectores $\mathbf{z}_i$, consiguiendo una reducción mayor de la dimensionalidad del espacio de características. Esto se puede conseguir mediante una inspección visual de la matriz de mezcla, seleccionando posteriormente aquellas fuentes que contribuyan en mayor grado a la formación de los datos observados, o a través de una ordenación decreciente de varianza y posterior selección del los máximos, ya que la transformación de ICA contendrá un paso previo en el que se realize PCA para decorrelacionar a primer orden. En la aproximación de deflación, también se pueden obtener un número menor de fuentes sin más que iterar el proceso de obtención de fuentes independientes un número limitado de veces.

10.3. ICA aplicado a imágenes representativas

Vimos en la sección 9.4 cómo pueden definirse K vectores representativos que sirvan para describir las imágenes de la base de datos a partir de su semejanza a éstos. Vimos que, tomado de esta manera, la información contenida en los vectores representativos estará fuertemente correlacionada, debido principalmente a que las diferencias entre estos vectores característicos son sutiles. Este hecho justificaba el uso de PCA, no como técnica de reducción dimensional, sino como método para obtener un nuevo conjunto

de K vectores de-correlacionados. Según la ecuación (9.15), la transformación de PCA aplicada a $\mathbf{R}$, definido previamente como el conjunto de vectores representativos, conduciría a un nuevo conjunto de K variables o componentes principales, cuya matriz de covarianza sería diagonal, es decir, estarían decorrelacionadas. En la práctica, los vectores representativos $\mathbf{\Psi}_k$ son estimados por las medias muestrales intra-clase $\mathbf{r}_k \in \mathbb{R}^m$, definidas como:

$$\mathbf{r}_k = \frac{1}{N_k} \sum_{\{y_i \in \mathcal{C}^k\}} \mathbf{\Gamma}_i \ , \qquad k = 1, 2, ..., K \tag{10.22}$$

donde N_k denota el número de imágenes en la clase $\mathcal{C}^k$.

Por las mismas razones, se puede usar ICA en este contexto para obtener un conjunto de K fuentes a partir de los vectores representativos, con la ventaja de que ICA proporcionará un conjunto de fuentes que no solo estarán decorrelacionadas a primer orden, sino que serán estadísticamente independientes. ICA habrá de aplicarse por tanto al conjunto de imagenes representativas $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_1, ..., \mathbf{r}_K]$ de manera que:

$$\mathbf{x}_i = \mathbf{w}_i^T \mathbf{R} \tag{10.23}$$

donde $\mathbf{x}_i$ son las $i = 1, ..., K$ son las fuentes independientes (ver figura 10.1) y la matriz de mezcla esta formada por los vectores $\mathbf{w}_i \in \mathbb{R}^m$ los $i = 1, ..., K$. Por lo tanto, se estimará un número K de fuentes independientes, es decir, el número de vectores representativos. En nuestros experimentos este valor tomará valores entre 2 y 4. Al igual que hicimos en el caso de PCA (9.38), podremos usar las fuentes independientes obtenidas como los K primeros vectores de una nueva base D de $\mathbb{R}^m$:

$$C = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K, \mathbf{u}_{K+1}, ..., \mathbf{u}_m\} \tag{10.24}$$

para describir el resto de la base de datos. Ahora, los vectores $\mathbf{x}_i$ tendrán la propiedad de ser estadísticamente independientes. Las primeras K coordenadas $\mathbf{z}_i = (z_1, z_2, ..., z_K)_i$ de un vector i de la base de datos D se usarán para clasificación, y vendrán dadas por:

$$\mathbf{z}_i = \mathbf{\Gamma}_i^T \mathbf{X} \quad i = 1, ..., n \tag{10.25}$$

con $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K]$. Esto implica que si el vector Γ_a guarde semejanza con alguno de los vectores representativos $\mathbf{x}_k$, sus coordenadas $\mathbf{z}_a$ en la base D serán, en principio, estadísticamente independientes de las coordenadas $\mathbf{z}_b$ del vector Γ_b semejante a *otro* vector representativo $\mathbf{x}_j$. Con esto se conseguirá, potencialmente, una separación mejor entre clases. ICA sirve en este contexto para obtener las características representativas de cada clase, independientes de las demás, sin necesidad de introducir ninguna información específica a cerca de la enfermedad. En este caso, se tienen en cuenta *todas* las fuentes obtenidas, y la reducción de la alta dimensionalidad del espacio de características se consigue a través del método de proyección definido a través de una ecuación del tipo (4.26), pasando de $\mathbb{R}^m$ a R^K, siendo K el número de vectores representativos.

10.4. Experimentos

Realizaremos 2 experimentos diferentes con ICA, cada uno con una base de datos diferente: el primero hará uso la base de datos de SPECT del hospital Virgen de las Nieves (ver sección 6.2, mientras que usarmos la base de datos de ADNI para el segundo de ellos (ver sección 6.1).

10.4.1. Experimentos con VN

Con la base de datos de VN se tomaron dos métodos para construir las imágenes representativas de (10.22):

- *Método I:* Usando 2 imágenes representativas; correspondientes a la clase NOR y ATD, combinando las ATD-1, ATD-2 y ATD-3 en una sola.
- *Método II:* Using 4 imágenes representativas; correspondientes a la clase NOR, ATD-1, ATD-2 and ATD-3. (ver 10.1)

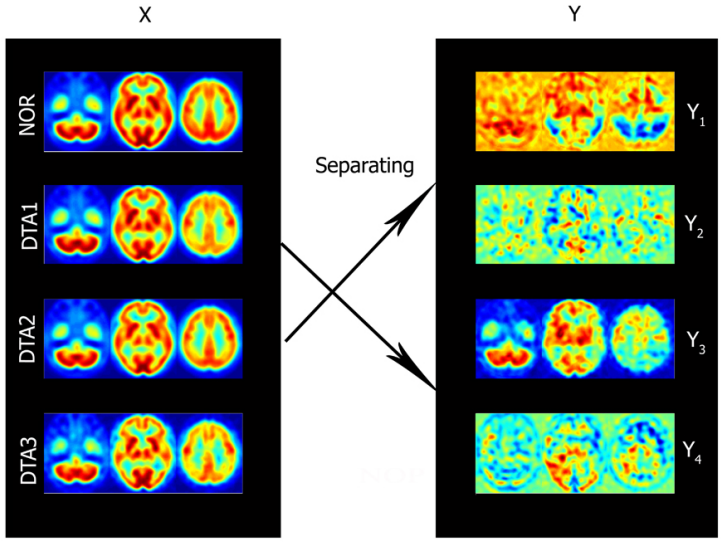

Figura 10.1: Proceso de separación de las 4 imagenes representativas originales en 4 fuentes independientes. Se muestran tres cortes transaxiales del cerebro.

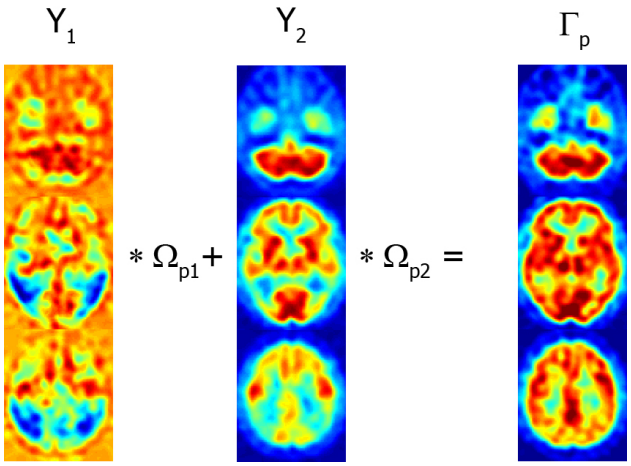

Figura 10.2: Representación de un paciente de Alzheimer en términos de los coeficientes de ICA siguiendo el método I.

Una vez obtenidos los vectores de entrenamiento de (4.26), se usaron para entrenar un SVM con 4 kernels diferentes: lineal, cuadratico, Radial Basis Function (RBF) and polinomial, y se testo usando una validación cruzada con la estrategia de dejar uno fuera. Hay que recalcar que aunque se usen más de dos clases para construir las imágenes representativas, los labels correspondientes a cada imágen serán solo de dos tipos ± 1. Así, todas las imágenes consideradas como ATD tendrán el mismo label, mientras que las Normales tendrán el contrario.

10.4.2. Experimentos con ADNI

En este experimento se sigue el método de ICA para obtener decorrelación entre imágenes representativas. Para obtener las imágenes representativas muestrales a partir de (10.22), se dividirán los datos en tres grupos diferentes. Cada grupo dará lugar a la construcción de un número diferente de imágenes representativas:

- **Group 1** Se consideran todas las imágenes de la base de datos. Se forman 3 imágenes representativas, una correspondiente a NC, otra a MCI y otra a AD.

- **Group 2** Se considera un subconjunto de los datos. No se tienen en cuenta los sujetos MCI, por lo que se forman 2 imágenes representativas AD y NC.
- **Group 3** e considera un subconjunto de los datos. No se tienen en cuenta los sujetos AD, por lo que se forman 2 imágenes representativas MCI y NC.

Una vez obtenidas las imágenes representativas en cada caso, se aplica la tranformación de ICA (10.23) para extraer los vectores de características. Se realizan también pruebas sin dar este último paso, con el fín de probar la necesidad de aplicar un método de decorrelación. Los vectores de características son usados para entrenar un clasificador SVM usando 2 kernels diferentes: uno lineal, y otro no-lineal (RBF). El test se embebe dentro de un proceso de validación cruzada basado en la técnica de dejar uno fuera, con el fín de estimar los parametros de eficiencia de la clasificación.

10.5. Resultados

10.5.1. Resultados con VN

Los resultados que se resumen en la tabla 10.1 ponen de manifiesto que la idea de encontrar algunas imágenes representativas para la caracterización de la EA es adecuada. Los dos métodos descritos de construcción de la matriz (Ω) consucen a una mejora significativa con respecto a otros sistemas CAD, como nuestro baseline (VAF)(Stoeckel et˜al., 2001), cuyos resultados se muestran como valores de referencia. Como era de esperar por razones teóricas (Vapnik, 1998), los kernels lineales generalizarán mejor que otros kernels cuando la dimensión del espacio de características es pequeña. Tanto el método I como el II representan una alta compresión de la gran cantidad de datos contenidas en las imágenes del cerebro a un pequeño número de características, 2 o 4 valores, respectivamente. El método I exhibe resultados interesantes, como una alta especificidad, alcanzando valores del $95,12\,\%$. Un gran número de estudios sobre los patrones típicos de perfusión del Alzheimer (Goethals et˜al., 2002; Kogure et˜al., 2000; Braak and Braak, 1997), muestran que las áreas cerebrales afectadas en la enfermedad del Alzheimer pueden

alcanzar diferentes niveles de hipo-perfusión a través de las distintas etapas. Esto explica el aumento de la sensibilidad del método II, lo que demuestra ser más adecuado para caracterizar la AD. Se obtiene mejor rendimiento con este segundo método cuando se combina con un núcleo rbf, llegando a $91,1\,\%$ precisión. Especificidad tomó el valor $92,7\,\%$ y la sensibilidad $89,5\,\%$ en ese caso.

Tabla 10.1: Medidas de precisión del método de ICA

	Parametro (%)	Kernel				VAF
		Lineal	Cuadratico	RBF	Polinomico	
Method I	Precisión	84.81	87.34	89.87	88.61	72.15
	Especificidad	87.80	92.68	95.12	92.68	78.05
	Sensibilidad	81.58	81.58	84.21	84.21	65.79
Method II	Precisión	83.54	86.07	91.14	88.60	74.68
	Epecificidad	85.37	87.80	92.68	90.24	82.93
	Sensibilidad	81.58	84.21	89.47	86.84	65.79

10.5.2. Resultados con ADNI

La mejora para la separabilidad de clases que produce ICA queda reflejada en la figura 10.3. Se puede llevar a cabo un análisis similar al de PCA, puesto que la componente independiente 1 (IC1) es muy parecida al Eigenbrain 2, aunque ahora la IC2 introduce cierta capacidad de discriminación, contrariamente a Eigenbrain 1. Este hecho se refleja en la pendiente del clasificador, que implica que una combinación de valores de $\mathbf{\Omega}_1$ y $\mathbf{\Omega}_2$ es necesaria para determinar la clase de cada imagen. La diferencia principal entre IC2 y el Eigenbrain 1 es que las regiones del cíngulo y temporo-paritales toman valores de intensidad extremadamente bajos, asi como también el cortex frontal toma valores de intensidad bajos. Esta última característica se asocia con los pacientes de Alzheimer con una afección severa, y por lo tanto implica que una combinación de valores bajos de $\mathbf{\Omega}_1$ y $\mathbf{\Omega}_2$ (la esquina inferior izquierda) representan sujetos severamente afectados por el Alzheimer, mientras que una combiandción de valores altos de $\mathbf{\Omega}_1$ y $\mathbf{\Omega}_2$ (la esquina superior derecha) representarán con mayor probabilidad patrones de imagenes de NC, siendo la región intermedia delimitada por el clasificador.

	Proyeccion		ICA		
	Kernel Lineal	kernel RBF	Kernel Lineal	kernel RBF	
Precicisión	76.26 %	76.26 %	88.13 %	89.50 %	Group 1
Especificidad	0 %	0 %	67.31 %	67.31 %	
Sensibilidad	100 %	100 %	94.61 %	96.41 %	
fpp	1	1	2.8940	2.9489	
fpn	∞	∞	0.0801	0.0534	
Precicisión	60.00 %	53.33 %	86.67 %	84.76 %	Group 2
Especificidad	65.38 %	65.38 %	88.46 %	84.62 %	
Sensibilidad	54.72 %	41.51 %	84.91 %	84.91 %	
fpp	1.5807	1.1992	7.3585	5.5189	
fpn	0.6926	0.8946	0.1706	0.1784	
Precicisión	68.67 %	68.67 %	81.33 %	83.73 %	Group 3
Especificidad	0 %	0 %	65.38 %	67.31 %	
Sensibilidad	100 %	100 %	88.60 %	91.23 %	
fpp	1	1	2.5595	2.7905	
fpn	∞	∞	0.1744	0.1303	

Tabla 10.2: Medidas del rendimiento del metodo de ICA para los tres grupos muestrales, en referencia a la simple proyección

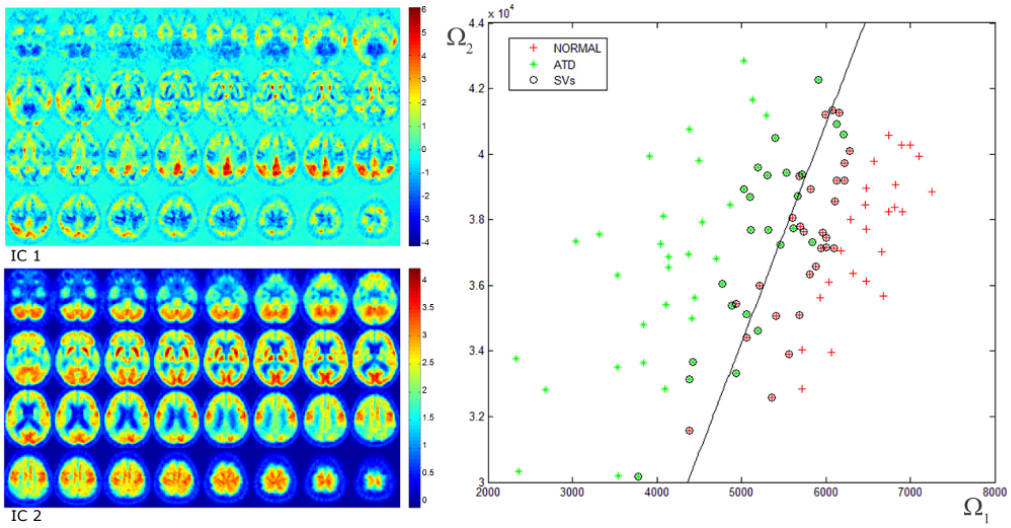

Figura 10.3: Clasificador lineal y vectores de entrenamiento $\mathbf{x}_i = [\mathbf{\Omega}_1, \mathbf{\Omega}_2]_i$ (derecha) obtenidos proyectando cada imagen en el espacio de ICs (izquierda, cortes transaxiales)

CAPÍTULO 11

Conclusiones

A modo de sumario, detallaremos a continuación las principales aportaciones y contribuciones contenidas en las diferentes partes de este trabajo

Parte I: Fundamentos En esta primera parte del trabajo se revisan los diferentes enfoques existentes para el diagnóstico del Alzheimer, a la vez que se introducen todas las herramientas que serán necesarias en posteriores análisis y desarrollos de la tesis. Se describe las características de la enfermedad del Alzheimer, así como las metodologías de diagnósitico y de monitorización mediante imágenes tomográficas funcionales, junto con las principales herramientas usadas para su diagnóstico, como SPM y otros sistemas CAD. Se introducen después los conceptos fundamentales de la teoría de aprendizaje estadístico con especial interés en el aprendizaje supervisado aplicado a problemas con pequeño número de muestras en comparación con la dimensión del espacio de características. Se desarrollan también las herramientas de análisis de rendimiento del aprendizaje estadístico mediante clasificación, como las curvas ROC

o los métodos de validación cruzada. Finalmente se destacan los clasificadores construidos a partir de Maquinas de Soporte Vectorial (SVM), que será la principal herramienta de clasificación usada posteriormente.

Parte II: Desarrollos Experimentales Se dedica una parte especial de este trabajo a la descripción minuciosa de los datos usados para desarrollar las técnicas de ayuda al diagnóstico. Se enfatiza el papel que juega el tratamiento previo que estos datos para la posterior tarea de clasificación. En este sentido, este trabajo aporta contribuciones novedosas, como la normalización en intensidad, a la vez que propone una metodología concreta para resolver el problema de comparación de imágenes de manera automática. Puesto que la tarea de clasificación se basa en la hipótesis de que diferentes vóxeles de una imágen estan en correspondencia espacial exacta con los diferentes vóxels de cualquier otra imagen, este preproceso que ha de llevarse a cabo toma un papel fundamental. En nuestro trabajo se encuentra una dependencia fuerte de los resultados de clasificación en este preprocesado, sobre todo en lo concerniente a la normalización en intensidad. Este hecho se refleja al usar los datos preprocesados por ADNI, siguiendo el método 4 descrito en 6.1, donde los resultados obtenidos están 10 puntos porcentuales por debajo de los resultados obtenidos mediante las técnicas desarrolladas por nosotros. Nuestra propuesta sencilla de normalización de intensidad juega con la ventaja de que tanto las imágenes SPECT como PET en enfermos de Alzheimer, muestran una *disminución* de la intensidad de regiones concretas del cerebro. Esto permite que la intensidad pueda ser normalizada a un máximo, sin que el patrón de hipoperfusión se vea distorsionado gravemente. Sin embargo este es un punto muy sutil, y necesita de un trabajo exhaustivo de análisis para desarrollar no solo una técnica de normalización sencilla, sino también la mas apropiada posible.

Parte III: Análisis en Componentes Esta parte describe las principales contribuciones de esta investigación, a través del desarrollo de tres técnicas CAD completamente nuevas: componentes, componentes principales, y componentes independientes. Las tres técnicas muestran un comportamiento general satisfactorio, con una precisión de clasificación alrededor del 90 %. La técnica de componentes destaca por su técnica de extracción de características, que es la más precisa de las tres.

La busqueda completa que se lleva a cabo haciendo uso de diferentes técnicas, como factorización y filtro de envoltura, consigue detectar de la manera mas exacta las regiones de interés para la clasificación. Una selección apropiada de los parámetros (tamaño y forma de las componentes) podría conducir a una precisión aún mayor de la presentada aqui. La contrapartida es que este proceso tiene un coste computacional alto, por lo que el fine-tunning de los parámetros es una tarea ardua.

Es destacable que el siguiente método, el más sencillo basado en PCA, consigue detectar regiones similares del cerebro como aquellas relevantes para clasificación, con un coste computacional significativamente menor. Se presentan 3 diferentes tecnicas de aplicación de PCA. Los resultados presentados animan a tomar el método de PCA combinado con la técnica de ordenación del Fisher Discriminant Ratio, como el mejor método presentado. Sin embargo, la estimación de la precisión de este método debe tomarse con cautela, debido principalmente al cálculo del factor FDR usando la base de datos completa podría estar sesgado. Por lo tanto, debería considerarse que la estimación de la precisión está sobrevaluado, siendo un valor más realista algo cercano al 90 %.

Es de esperar que la mejora real venga de la consideración de las relaciones entre voxels de orden superior, dada por ICA. Este método se presenta como una generalización de PCA al caso de independencia estadística. La novedosa técnica de aplicación de ICA basada en la construcción de imágenes representativas ha resultado superar la precisión del método de PCA, como puede verse en la tabla 11.1. Esta técnica contiene implícitamente un método de selección de componentes independientes, basado en la idea de representar las imágenes cerebrales en términos de un grupo reducido de imágenes estadísticamente independientes. El hecho de que la mejora no supere los 2 puntos porcentuales, sugiere que la distribución de probabilidad de las diferentes clases está cercana a la gaussianidad, hecho que podría ser utilizando en trabajos futuros.

Una comparación de los tres métodos elevaría al método de ICA basado en imágenes representativas como el más eficiente y el más eficaz. Sin embargo, aunque en el presente estadio sea así, nuestra creencia es que la eficiencia del método de las componentes esta estimada a la baja, ya que la validación cruzada de pliegues no proporciona una referencia

	PCA		ICA		
	Linear Kernel	RBF kernel	Linear Kernel	RBF kernel	
Accuracy	90.41 %	90.41 %	88.13 %	89.50 %	Group 1
Specificity	75.00 %	71.15 %	67.31 %	67.31 %	
Sensitivity	95.21 %	96.41 %	94.61 %	96.41 %	
Positive Likelihood	3.8084	3.3421	2.8940	2.9489	
Negative Likelihood	0.0639	0.0505	0.0801	0.0534	
Accuracy	84.76 %	84.76 %	86.67 %	84.76 %	Group 2
Specificity	86.54 %	86.54 %	88.46 %	84.62 %	
Sensitivity	83.02 %	83.02 %	84.91 %	84.91 %	
Positive Likelihood	6.1671	6.1671	7.3585	5.5189	
Negative Likelihood	0.1962	0.1962	0.1706	0.1784	
Accuracy	82.53 %	83.13 %	81.33 %	83.73 %	Group 3
Specificity	67.31 %	67.31 %	65.38 %	67.31 %	
Sensitivity	89.47 %	90.35 %	88.60 %	91.23 %	
Positive Likelihood	2.7368	2.7637	2.5595	2.7905	
Negative Likelihood	0.1564	0.1434	0.1744	0.1303	

Tabla 11.1: Comparación de las medidas de precisión de los métodos de ICA y PCA, para los 3 grupos.

exacta. Además, una adecuación de los parametros podría significar una mejora importante. Sin embargo, es necesario desarrollar técnicas computacionalmente más eficientes con este fin.

11.1. Trabajo futuro

Tras la investigación realizada y presentada en este trabajo, se abren las puertas a nuevas líneas de investigación basadas en los hallazgos de este trabajo. Principalmente se pueden resumir en tres grandes líneas:

Normalización en Intensidad La importancia de la normalización en intensidad puesta de manifiesto anteriormente sugiere la elaboración de técnias de normalización espacial mas refinadas. El siguiente paso lógico sería el uso de regiones del cerebro cuya intensidad sea invariante como referencia a la hora de normalización, en lugar de usar un máximo. Una de estas regiones la constituye el cerebelo. Para usar la intensidad media del cerebelo como referencia para normalizar el conjunto de imágenes es necesario elaborar una técnica de identificación automática de ésta region.

Una vez conseguido este objetivo, se abrirían las puertas al análisis de imágenes PET usando el trazador ^{11}C PIB, para las que ADNI proporcina una amplia variedad de imágenes. Además se tienen imágenes de los dos tipos de trazadores de un mismo paciente, lo que tambien permitiría hacer comparaciones entre las diferentes drogas disponibles para el analísis tomográfico. En las figuras 11.1 y 11.2, se presentan unos resultados preliminares tomados con los mismos 36 pacientes, y normalizados en intensidad a un máximo. Este analisis debe extenderse aun grupo mas amplio, debido a la prevalencia de la clase del Alzheimer en este subgrupo y debido a las limitaciones de la normalización a un máximo en el caso del ^{11}C PIB.

Automatización de selección de componentes El método de componentes ha sido presentado como una técnica de alta precisión en la detección de reigones de interés. Sin embargo, en este estadio presenta unas limitaciones computacionales grandes. El siguiente paso en desarrollo es la elaboración de técnicas automaticas de selección de componentes, sin

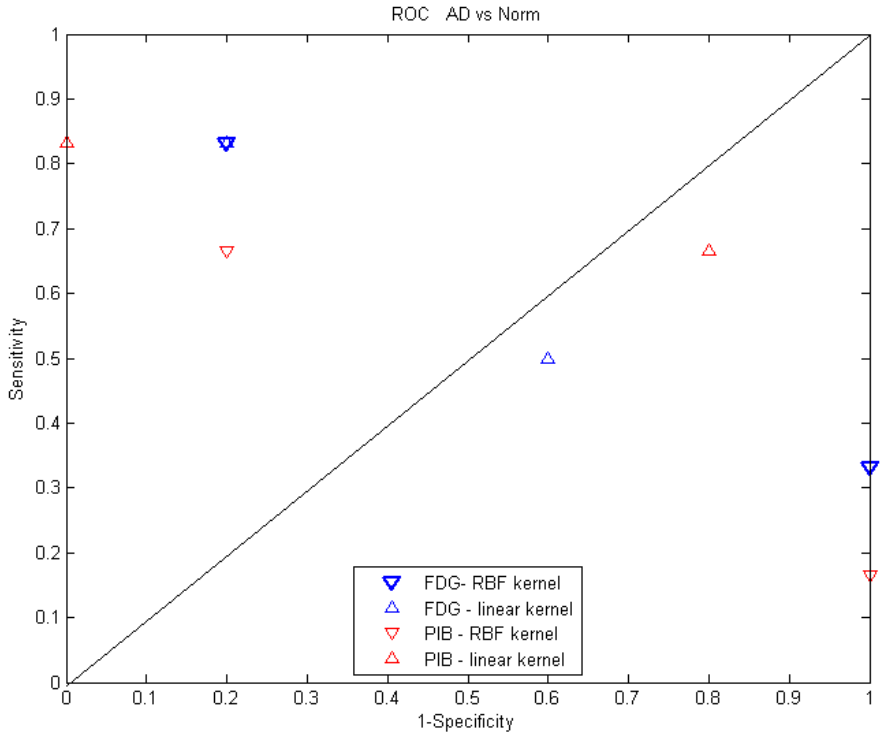

Figura 11.1: Representación en el espacio ROC de los resultados de clasificación de los mismos pacientes usando las técnicas de proyección, PCA e ICA para diferentes trazadores y kernels. Se tienen en cuenta solo los pacientes AD y los NC

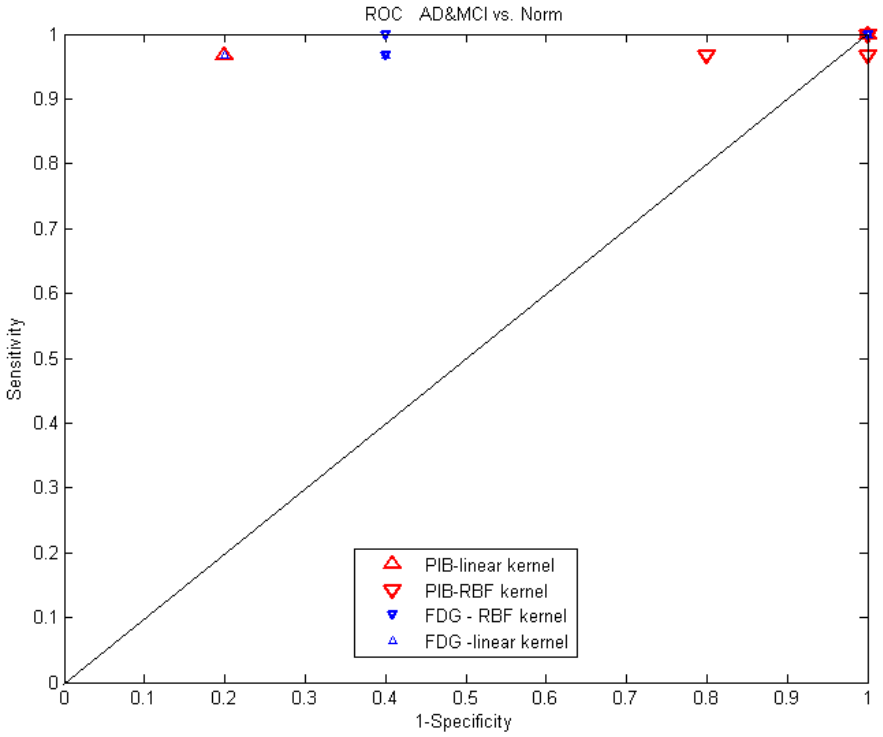

Figura 11.2: Representación en el espacio ROC de los resultados de clasificación de los mismos pacientes usando las técnicas de proyección, PCA e ICA para diferentes trazadores y kernels. Se tienen en cuenta todos los pacientes AD, MCI y NC

necesidad de completar un filtro de envoltura sobre la imagen completa. Este esta basado en una técnica hibrida en desarrollo, que combina elementos de algoritmos genéticos asi como la técnica ADAboost de selección de características. En líneas genereales se buscan componentes elegidas al azar que consituyan o esten cercanas a regiones de interés. Tras una evaluación se asigna un peso a las regiones estudiadas y a sus alrededores, haciendo que el método converga automaticamente a aquellas regiones de máxima precisión en un tiempo mucho menor.

ICA como LDA Linear Discriminant Analisis (LDA) es una técnica que combina los conceptos de PCA y el criterio de Fisher para seleccionar componentes, haciendo PCA no sobre el conjunto de datos, sino de forma intra-clase. Nuestra métodología de ICA aplicada a imágenes representativas guarda relación con estos conceptos, lo que sugiere un paso siguiente en la dirección de la contrucción de una técnica similar a LDA usando ICA.

Bibliografía

A., D. P., Kittler, J. (Eds.), 1982. Pattern Recognition: A Statistical Approach. Prentice-Hall, London.

Adler, R., 1981. The Geometry of random fields. Wiley, New York.

Alexander, G. E., Chen, K., Pietrini, P., Rapoport, S. I., Reiman, E. M., May 2002. Longitudinal PET evaluation of cerebral metabolic decline in dementia: A potential outcome measure in alzheimer's disease treatment studies. The American Journal of Psychiatry 159 (5), 738–45, PMID: 11986126.

Ashburner, J., Friston, K. J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7 (4), 254–66.

Ayache, N., 1996. Analyzing 3d images of the brain. NeuroImage 4 (3), S34–S35.

Bell, A. J., Sejnowski, T. J., Nov 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7 (6), 1129–1159.

Bingham, E., 2003. Advances in independent component analysis with applications to data mining. Ph.D. thesis, Helsinki University of Technology.

Blum, A. L., Langley, P., 1997. Selection of relevant features and examples in machine learning. ARTIFICIAL INTELLIGENCE 97, 245—271.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.4823

Braak, H., Braak, E., 1997. diagnostic criteria for neuropathologic assessment of alzheimer's disease. Neurobiology and Aging 18 (4), S85–S88.

Breiman, L., 1999. Pasting small votes for classification in large database and on-line. Machine Learning 36, 85–103.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R., 1984. Classification and Regression Trees, 1st Edition. Chapman & Hall/CRC.

Bruyant, P. P., 2002. Analytic and iterative reconstruction algorithms in spect. The Journal of Nuclear Medicine 43 (10), 1343–1358.

Burges, C. J. C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Cardoso, J., 1990. Eigen-Structure of the Fourth-Order cumulant tensor with application to the blind source separation problem. In: Proceedings of ICASSP.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.3319

Cardoso, J., Souloumiac, A., 1993. Blind beamforming for non gaussian signals. IEE PROCEEDINGS-F 140, 362—370.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5684

Carr, D. B., Goate, A., Phil, D., Morris, J. C., Sep. 1997. Current concepts in the pathogenesis of alzheimer's disease. The American Journal of Medicine 103 (3A), 3S–10S, PMID: 9344401.
URL http://www.ncbi.nlm.nih.gov/pubmed/9344401

Chase, T. N., Foster, N. L., Fedio, P., Brooks, R., Mansi, L., Chiro, G. D., 1984. Regional cortical dysfunction in alzheimer's disease as determined by positron emission tomography. Annals of Neurology 15 Suppl, S170–4, PMID: 6611118.

Chornoboy, E. S., Chen, C. J., Miller, M. I., Miller, T. R., Snyder, D. L., 1990. An evaluation of maximum likelihood reconstruction for spect. IEEE Transactions on Medical Imaging 9 (1), 99–110.

Claus, J. J., van Harskamp, F., Breteler, M. M. B., Krenning, E. P., de Koning abd J. M. van der Cammen, I., Hofman, A., Hasan, D., 1994. The diagnostic value of SPECT with tc 99m HMPAO in alzheimer's disease. a population-based study. Neurology 44 (3), 454–461.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36 (3), 287–314.

Cover, T. M., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers 14 (3), 326–334.

Cummings, J. L., Vinters, H. V., Cole, G. M., Khachaturian, Z. S., 1998. Alzheimer's disease: etiologies, pathophysiology, cognitive reserve, and treatment opportunities. Neurology 51 (suppl. 1), S2–S17.

de Leon, M. J., Ferris, S. H., George, A. E., Reisberg, B., Christman, D. R., Kricheff, I. I., Wolf, A. P., Sep. 1983. Computed tomography and positron emission transaxial tomography evaluations of normal aging and alzheimer's disease. Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism 3 (3), 391–4, PMID: 6603463.

Devijver, P. A., Kittler, J., 1982. Pattern Recognition: A Statistical Approach, first edition Edition. Prentice Hall.

Doak, J., 1992. An evaluation of feature-selection methods and their application to computer security. Tech. rep., University of California, Department of Computer Science.

Drzezga, A., Lautenschlager, N., Siebner, H., Riemenschneider, M., Willoch, F., Minoshima, S., Schwaiger, M., Kurz, A., Aug. 2003. Cerebral metabolic changes accompanying conversion of mild cognitive impairment into alzheimer's disease: a PET follow-up study. European Journal of Nuclear Medicine and Molecular Imaging 30 (8), 1104–13, PMID: 12764551.

Duara, R., Grady, C., Haxby, J., Sundaram, M., Cutler, N. R., Heston, L., Moore, A., Schlageter, N., Larson, S., Rapoport, S. I., Jul. 1986. Positron emission tomography in alzheimer's disease. Neurology 36 (7), 879.

Duin, R. P. W., 2000. Classifiers in almost empty spaces. In: Proceedings 15th International Conference on Pattern Recognition. Vol. 2. IEEE, pp. 1–7.

English, R. J., Childs, J. (Eds.), 1996. SPECT: Single-Photon Emission Computed Tomography: A Primer. Society of Nuclear Medicine.

Enqing, D., Guizhong, L., Yatong, Z., Xiaodi, Z., 2002a. Applying support vector machines to voice activity detection. In: 6th International Conference on Signal Processing. Vol. 2. pp. 1124–1127.

Enqing, D., Heming, Z., Yongli, L., 2002b. Low bit and variable rate speech coding using local cosine transform. In: Proc. of the 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON '02). Vol. 1. pp. 423–426.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Annals Eugen. 7, 188, 179.

Foster, N. L., Chase, T. N., Fedio, P., Patronas, N. J., Brooks, R. A., Chiro, G. D., Aug. 1983. Alzheimer's disease: Focal cortical changes shown by positron emission tomography. Neurology 33 (8), 961.

Foster, N. L., Chase, T. N., Mansi, L., Brooks, R., Fedio, P., Patronas, N. J., Chiro, G. D., Dec. 1984. Cortical abnormalities in alzheimer's disease. Annals of Neurology 16 (6), 649–54, PMID: 6335378.

Frackowiak, R. S. J., Ashburner, J. T., Penny, W. D., Zeki, S., December 2003. Human Brain Function, Second Edition. Academic Press.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W., 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press.

Fukunaga, K., Sep. 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press, New York.

Fung, G., Stoeckel, J., 2007. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. Knowledge and Information Systems 11 (2), 243–258.

Goethals, I., van deWiele, C., Slosman, D., Dierckx, R., 2002. Brain SPECT perfusion in early Alzheimer disease: where to look? European Journal of Nuclear Medicine 29 (8), 975–978.

Górriz, J. M., Ramírez, J., Lassl, A., Salas-Gonzalez, D., Lang, E. W., Puntonet, C. G., Álvarez, I., López, M., Gómez-Río, M., 2008. Automatic computer aided diagnosis tool using component-based svm. In: 2008 IEEE Nuclear Science Symposium Conference Record. pp. 4392–4395.

Hellman, R. S., Tikofsky, R. S., Collier, B. D., Hoffmann, R. G., Palmer, D. W., Glatt, S., Antuono, P. G., Isitman, A. T., Papke, R. A., 1989. Alzheimer disease: quantitative analysis of i-123-iodoamphetamine spect brain imaging. Radiology 172, 183–188.

Higdon, R., Foster, N. L., Koeppe, R. A., DeCarli, C. S., Jagust, W. J., Clark, C. M., Barbas, N. R., Arnold, S. E., Turner, R. S., Heidebrink, J. L., Minoshima, S., 2004. A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging. Statistics in Medicine 23, 315–326.

Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., Crain, B., Hulette, C., Earl, N., Coleman, R. E., Nov. 2000. FDG PET imaging in patients with pathologically verified dementia. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine 41 (11), 1920–8, PMID: 11079505.

Holman, B. L., Johnson, K. A., Gerada, B., Carvalho, P. A., Satlin, A., 1992. The scintigraphic appearance of alzheimer's disease: A prospective study using technetium-99m-hmpao spect. J. Nucl. Med. 33 (2), 181–185.

Hudson, H. M., Larkin, R. S., 1994. Accelerated image reconstruction using ordered subsets of projection data. IEEE Transactions on Medical Imaging 13 (4), 601–609.

Hyvaerinen, A., 1999. Fast and robust Fixed-Point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10 (3), 634, 626.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.4731

Ibañez, V., Pietrini, P., Alexander, G. E., Furey, M. L., Teichberg, D., Rajapakse, J. C., Rapoport, S. I., Schapiro, M. B., Horwitz, B., Jun. 1998. Regional glucose metabolic abnormalities are not the result of atrophy in alzheimer's disease. Neurology 50 (6), 1585–93, PMID: 9633698.

Illán, I. A., Górriz, J. M., Ramírez, J., Salas-González, D., López, M., Puntonet, C. G., Segovia, F., 2009. Alzheimer's diagnosis using eigenbrains and support vector machines. IET Electronics Letters 45 (7), 342–343.

Ishii, K., Kono, A. K., Sasaki, H., Miyamoto, N., Fukuda, T., Sakamoto, S., Mori, E., 2006. Fully automatic diagnostic system for early- and late-onset mild Alzheimer's disease using FDG PET and 3D-SSP. European Journal of Nuclear Medicine and Molecular Imaging 33 (5), 575–583.

J., T., P., T., 1988. A Co-planar Stereotatic Atlas of the Human Brain. Stuttgart: Thieme.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In: Lecture Notes in Computer Science. Vol. 1398. pp. 137–142.

John, G. H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: International Conference on Machine Learning. pp. 121–129, journal version in AIJ, available at http://citeseer.nj.nec.com/13663.html.
URL http://citeseer.ist.psu.edu/john94irrelevant.html

Johnson, K. A., Kijewski, M. F., Becker, J. A., Garada, B., Satlin, A., Holman, B. L., 1993. Quantitative brain spect in alzheimer's disease and normal aging. J. Nucl. Med. 34 (11), 2044–2048.

Jolliffe, I. T., 2002. Principal Component Analysis, 2nd Edition. Springer series in statistics. Springer.

Kalatzis, I., Pappas, D., Piliouras, N., Cavouras, D., 2003. Support vector machines based analysis of brain SPECT images for determining cerebral abnormalities in asymptomatic diabetic patients. Medical Informatics and the Internet in Medicine 28 (3), 221–230.

Karas, G., Scheltens, P., Rombouts, S., Visser, P., van Schijndel, R., Fox, N., Barkhof, F., Oct. 2004. Global and local gray matter loss in mild cognitive impairment and alzheimer's disease. NeuroImage 23 (2), 708–716.

Kim, K. I., Jung, K., Park, S. H., Kim, H. J., 2002. Support vector machines for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (11), 1542–1550.

Kogure, D., Matsuda, H., Ohnishi, T., Asada, T., Uno, M., Kunihiro, T., Nakano, S., Takasaki, M., 2000. Longitudinal evaluation of early Alzheimer disease using brain perfusion SPECT. The Journal of Nuclear Medicine 41 (7), 1155–1162.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Vol. 2. pp. 1137–1143.

Kohavi, R., John, G. H., 1997. Wrappers for feature subset selection. Artificial Intelligence - Special issue on relevance 97 (1-2), 273–324.

Langbaum, J. B., Chen, K., Lee, W., Reschke, C., Bandy, D., Fleisher, A. S., Alexander, G. E., Foster, N. L., Weiner, M. W., Koeppe, R. A., Jagust, W. J., Reiman, E. M., May 2009. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the alzheimer's disease neuroimaging initiative (ADNI). NeuroImage 45 (4), 1107–1116.

Lange, K., Carson, R., 1984. Em reconstruction for emission and transmission tomography. Journal of Computer Assisted Tomography 8, 306–312.

López, M., Ramírez, J., Górriz, J. M., Salas-González, D., Illan, I. A., Segovia, F., Puntonet, C. G., 2009. Automatic tool for the alzheimer's disease diagnosis using pca and bayesian classification rules. IET Electronics Letters 45 (8), 389–391.

McGeer, E. G., Peppard, R. P., McGeer, P. L., Tuokko, H., Crockett, D., Parks, R., Akiyama, H., Calne, D. B., Beattie, B. L., Harrop, R., Feb. 1990. 18Fluorodeoxyglucose positron emission tomography studies in presumed alzheimer cases, including 13 serial scans. The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques 17 (1), 1–11, PMID: 2311010.

Messa, C., Perani, D., Lucignani, G., Zenorini, A., Zito, F., Rizzo, G., Grassi, F., Del Sole, A., Franceschi, M., Gilardi, M. C., Fazio, F., 1994. High-Resolution Technetium-99m-HMPAO SPECT in Patients with Probable Alzheimer's Disease: Comparison with Fluorine-18-FDG PET. J Nucl Med 35 (2), 210–216.

Minoshima, S., Foster, N., Kuhl, D., Sep. 1994. Posterior cingulate cortex in alzheimer's disease. The Lancet 344 (8926), 895.

Minoshima, S., Frey, K. A., Koeppe, R. A., Foster, N. L., Kuhl, D. E., Jul. 1995. A diagnostic approach in alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine 36 (7), 1238–48, PMID: 7790950.

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L., Kuhl, D. E., Jul. 1997. Metabolic reduction in the posterior cingulate cortex in very early alzheimer's disease. Annals of Neurology 42 (1), 85–94, PMID: 9225689.

Miranda, A., Borgne, Y. L., Bontempi, G., Jun. 2008. New routes from minimal approximation error to principal components. Neural Processing Letters 27 (3), 197–207.
URL `http://dx.doi.org/10.1007/s11063-007-9069-2`

Morris, J., 1993. Clinical dementia rating. Neurology 43, 2412–2414.

Mosconi, L., Tsui, W. H., Herholz, K., Pupi, A., Drzezga, A., Lucignani, G., Reiman, E. M., Holthoff, V., Kalbe, E., Sorbi, S., Diehl-Schmid, J., Perneczky, R., Clerici, F., Caselli, R., Beuthien-Baumann, B., Kurz, A., Minoshima, S., de Leon, M. J., Mar. 2008. Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, alzheimer's disease, and other dementias. J Nucl Med 49 (3), 390–398.

Nitrini, R., Buchpiguel, C., Caramelli, P., Bahia, V., Mathias, S., Nascimento, C., Degenszajn, J., Caixeta, L., 2000. Spect in alzheimer's disease: features associated with bilateral parietotemporal hypoperfusion. Acta Neurologia Scandinava 101 (3), 172–176.

Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483–1492.

por Aapo Hyvaerinen, E., Karhunen, J., Oja, E., 2001. Independent component analysis. Wiley, New York.

Qi, F., Bao, C., Liu, Y., 2004. A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. In: International Symposium on Chinese Spoken Language Processing. pp. 77–80.

Ramírez, J., Górriz, J. M., Gómez-Río, M., Romero, A., Chaves, R., Lassl, A., Rodríguez, A., Puntonet, C. G., Theis, F., Lang, E., 2008. Effective emission tomography image reconstruction algorithms for SPECT data. Lecture Notes in Computer Science 5101, 741–748.

Ramírez, J., Yélamos, P., Górriz, J. M., Puntonet, C. G., Segura, J. C., 2006a. SVM-enabled voice activity detection. In: Lecture Notes in Computer Science. Vol. 3972. pp. 676–681.

Ramírez, J., Yélamos, P., Górriz, J. M., Segura, J. C., 2006b. SVM-based speech endpoint detection using contextual speech features. Electronics Letters 42 (7), 877–879.

Raudys, S., Duin, R. P. W., April 1998. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recognition Letters 19 (5-6), 385–392.

Raudys, S., Jain, A., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13 (3), 252–264.

Salas-González, D., Górriz, J. M., Ramírez, J., Lassl, A., Puntonet, C. G., 2008. Improved gauss-newton optimization methods in affine registration of spect brain images. IET Electronics Letters 44 (22), 1291–1292.

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., Lassl, A., Puntonet, C. G., Lang, E. W., Gómez-Río, M., 2008. A comparison of nonlinear least-square optimization methods in affine registration of spect images. In: 2008 IEEE Nuclear Science Symposium Conference Record. pp. 4396–4398.

Saxena, P., Pavel, D. G., Quintana, J. C., Horwitz, B., 1998. An automatic thresholdbased scaling method for enhancing the usefulness of Tc-HMPAO SPECT in the diagnosis of Alzheimers disease. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI, Lecture Notes in Computer Science. Vol. 1496. pp. 623–630.

Sheikh, J., Yesavage, J., 1986. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. NY: The Haworth Press.

Silverman, D. H., Small, G. W., Chang, C. Y., 2001. Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome. Journal of the American Medical Association 286 (17), 2120–2127.

Stoeckel, J., Ayache, N., Malandain, G., Koulibaly, P. M., Ebmeier, K. P., Darcourt, J., 2004. Automatic classification of spect images of alzheimer's disease patients and control subjects. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. Vol. 3217 of Lecture Notes in Computer Science. Springer, pp. 654–662.

Stoeckel, J., Malandain, G., Migneco, O., Koulibaly, P. M., Robert, P., Ayache, N., Darcourt, J., 2001. Classification of spect images of normal subjects versus images of alzheimer's disease patients. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. Vol. 2208 of Lecture Notes in Computer Science. Springer, pp. 666–674.

Talbot, P. R., Lloyd, J. J., Snowden, J. S., Neary, D., Testa, H. J., 1998. A clinical role for 99mTc-HMPAO SPECT in the investigation of dementia? J Neurol Neurosurg Psychiatry 64 (3), 306–313.

Tao, D., Tang, X., Li, X., Wu, X., 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7), 1088–1099.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of congnitive neuroscience 3 (1), 71–86.

Vapnik, V. N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.

Vapnik, V. N., 1998. Statistical Learning Theory. John Wiley and Sons, Inc., New York.

Vardi, Y., Shepp, L. A., Kaufman, L., 1985. A statistical model for positron emission tomography. Journal of the American Statistical Association 80 (389), 8–20.

Woods, R. P., 2000. Spatial transformation models. In: Bankman, I. N. (Ed.), Handbook of Medical Imaging. Academic Press, San Diego, Ch. 29, pp. 465–490.

Yélamos, P., Ramírez, J., Górriz, J. M., Puntonet, C. G., Segura, J. C., 2006. Speech event detection using support vector machines. In: Lecture Notes in Computer Science. Vol. 3991. pp. 356–363.

Ziolko, S. K., Weissfeld, L. A., Klunk, W. E., Mathis, C. A., Hoge, J. A., Lopresti, B. J., DeKosky, S. T., , Price, J. C., 2006. Evaluation of voxel-based methods for the statistical analysis of pib pet amyloid imaging studies in alzheimer's disease. Neuroimage 33 (1), 94–102.